

UCLA

UCLA Electronic Theses and Dissertations

Title

Planning Experiments with Causal Graphs

Permalink

<https://escholarship.org/uc/item/3b629605>

Author

Matiasz, Nicholas John

Publication Date

2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Planning Experiments with Causal Graphs

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Bioengineering

by

Nicholas John Matiasz

2018

© Copyright by
Nicholas John Matiasz
2018

ABSTRACT OF THE DISSERTATION

Planning Experiments with Causal Graphs

by

Nicholas John Matiasz

Doctor of Philosophy in Bioengineering

University of California, Los Angeles, 2018

Professor Alex Anh-Tuan Bui, Co-Chair

Professor Alcino Jose Silva, Co-Chair

Scientists aim to design experiments and analyze evidence to obtain maximum knowledge. Although scientists have many statistical methods to guide how they analyze evidence, they have relatively few methods to quantify the convergence of evidence, to explore the full range of consistent causal explanations, and to design subsequent experiments on the basis of such analyses. The goal of this research is to establish tools that use graphical models to perform causal reasoning and experiment planning. This dissertation presents and evaluates methods that allow scientists (1) to quantify both the convergence and consistency of evidence, (2) to identify every causal structure that is consistent with evidence reported in literature, and (3) to design experiments that can efficiently reduce the number of viable causal structures. This suite of methods is demonstrated with real examples drawn from neuroscience literature.

This dissertation shows how scientific results can be merged to yield new inferences by determining whether the results are consistent with various causal structures. Also presented is a Bayesian model of scientific consensus building, based on the principles of convergence and consistency. Together, these approaches form the basis of a mathematical framework that complements statistics: quantitative formalisms can be used not only to demonstrate each result's significance but also to justify each experiment's design.

The dissertation of Nicholas John Matiasz is approved.

Denise R. Aberle

William Hsu

Ricky Kiyotaka Taira

Alex Anh-Tuan Bui, Committee Co-Chair

Alcino Jose Silva, Committee Co-Chair

University of California, Los Angeles

2018

When scientists seek to learn new, interesting truths, to find important patterns hiding in vast arrays of data, they are often trying to do something like searching for a needle in a really huge haystack of falsehoods, for a correct network among many possible networks, for a robust pattern among many apparent but unreal patterns.

— CLARK GLYMOUR

Statisticians can no longer ignore the mental representation in which scientists store experiential knowledge, since it is this representation, and the language used to access it that determine the reliability of the judgments upon which the analysis so crucially depends.

— JUDEA PEARL

One may be tempted to assume that whenever we ask questions of nature, of the world there outside, there is reality existing independently of what can be said about it. We will now claim that such a position is void of any meaning. It is obvious that any property or feature of reality “out there” can only be based on information we receive. There cannot be any statement whatsoever about the world or about reality that is not based on such information. It therefore follows that the concept of a reality without at least the ability in principle to make statements about it to obtain information about its features is devoid of any possibility of confirmation or proof. This implies that the distinction between information, that is knowledge, and reality is devoid of any meaning.

— ANTON ZEILINGER

TABLE OF CONTENTS

1	Introduction	1
1.1	Motivation: The imprecision of biological pathway diagrams	3
1.2	Aims: Practical causal-reasoning tools for scientists	7
1.3	Overview: Planning experiments with causal graphs	8
2	Literature review	10
2.1	Research maps	10
2.1.1	Categorizing evidence	11
2.1.2	Quantifying evidence	13
2.2	Causal graphs	17
2.2.1	Markov equivalence classes	20
2.3	Causal discovery	21
2.4	Experiment selection	23
2.5	Gaps in the literature	24
2.5.1	Meta-analytic methods that quantify evidential convergence	24
2.5.2	Causal discovery without primary data	27
2.5.3	Interpretable experiment-selection strategies	28
2.6	Contributions of this dissertation	30
2.6.1	The cumulative evidence index (CEI)	30
2.6.2	A literature-based technique for causal discovery	30
2.6.3	Interpretable heuristics for experiment selection	32
3	ResearchMaps: a web application for experiment planning	33
3.1	Implementation of the research-map framework	33
3.2	Creating research maps: The local map	35
3.3	Querying research maps: The global map	37
3.4	Data collected for analysis	38
3.5	Details of the software implementation	38

4	Collecting constraints on causal structure	44
4.1	Annotating empirical results in literature	45
5	Identifying consistent causal structures	47
6	Quantifying evidence and causal underdetermination	52
6.1	Quantifying evidence in research maps	52
6.2	Quantifying causal underdetermination in causal graphs	57
7	Selecting the next experiment	61
7.1	Maximizing evidence in research maps	61
7.2	Minimizing underdetermination in causal graphs	63
8	Evaluation	71
8.1	Reasoning with structural patterns in research maps	71
8.2	Reasoning with degrees of freedom in causal graphs	73
8.3	Simulations of experiment-selection algorithms	78
9	Conclusion	85
9.1	Summary of contributions	85
9.2	Assessment of hypotheses	86
9.3	Generalizability of the results	87
9.4	Range of applicability	88
9.5	Future work	89
9.5.1	Automating literature annotation	89
9.5.2	Extending the research-map schema	90
9.5.3	Generalizing the cumulative evidence index in research maps	91
9.5.4	Scaling SAT-based causal discovery methods	91
9.5.5	Incorporating sign information into causal discovery	92
9.5.6	Improving experiment-selection heuristics	92
	References	93

LIST OF FIGURES

1.1	A depiction of the scientific process from the perspective of basic-science researchers who perform experiments (e.g., molecular biologists). See Clark and Kinoshita [CK07] and Russ et al. [RRH11] for alternative depictions.	2
1.2	An example of a pathway diagram that has been adapted from a research article [CS03]. This diagram illustrates biological mechanisms, but because the meaning of each edge is not precisely defined, this diagram cannot necessarily be used to reason causally about the system.	4
1.3	The union of pathway diagrams is not necessarily consistent with the evidence that is encoded in—or with the inferences that are allowed by—the individual pathway diagrams. Note that the $X \rightarrow Z$ edge implies that X can affect Z independently of Y , even though the evidence that gave rise to the original pathways does not guarantee that this interpretation is correct.	6
1.4	This system diagram gives an overview of a meta-analytic approach to causal discovery and experiment selection. (In the research map, the studies involving X and Z are shown on separate edges to highlight their correspondence to the third and fourth statistical relations.)	9

2.1	This research map represents the empirical results and hypothetical assertions reported in a neuroscience article [CFK02]. All three types of relations are shown: for instance, an excitatory edge from K-ras to LTP, an inhibitory edge from NF1 to GABA inhibition, and a no-connection edge from N-ras to hippocampal learning. The symbol on the edge from NF1 to hippocampal learning (\downarrow) indicates that at least one negative intervention was performed to test the relation between these two phenomena. The edges in gray—from GABA inhibition to LTP, and from LTP to hippocampal learning—are hypothetical edges: putative causal assertions that lack empirical evidence. Hypothetical edges are useful for incorporating assumptions or background knowledge about a causal system; they give the research map additional structure to facilitate the interpretation of empirical results. © 2017 Matiasz et al. [MWW17a]; licensed under CC BY 4.0.	13
2.2	The heuristic approach for calculating an evidence score, which was used in early versions of research maps. Each cell in the table starts with a value of zero, and each empirical result increments the appropriate cell by one. The function $\text{Max}(\mathcal{E}, \mathcal{N}, \mathcal{I})$ returns the maximum value from the set $\{\mathcal{E}, \mathcal{N}, \mathcal{I}\}$. The edge's relation is assigned according to this maximum value: either excitatory (\mathcal{E}), no-connection (\mathcal{N}), or inhibitory (\mathcal{I}). © 2018 Matiasz et al. [MWD18]; licensed under CC BY 4.0.	16
2.3	The three causal graphs on the left form a Markov equivalence class. Although their edges have different orientations, these three causal graphs all imply the (in)dependence relations on the right, in accordance with the rules of d-separation.	22

3.1	A research map of a published article [CAS16]. Each node in a research map has three properties: what (top), where (middle), and when (bottom). Nodes are connected by edges that represent relations: excitatory (sharp arrowhead), inhibitory (blunt arrowhead), and no-connection (dotted line, circular arrowhead). Each empirical edge also has a CEI that reflects the amount of evidence represented, as well as symbols that reflect the study classes recorded for that edge. CEIs and study symbols are not assigned to hypothetical edges. Users can highlight edges that reflect the article’s main ideas, so that they are more apparent. In cases where no one relation has received dominant evidence, the corresponding edge is represented by a diamond arrowhead and is not assigned a CEI. © 2018 Matiasz et al. [MWD18]; licensed under CC BY 4.0.	35
3.2	Using hypothetical edges to organize research maps. The diagram above shows how hypothetical edges (in gray) help to organize empirical edges in a research map, thus framing the empirical results in light of a specific hypothesis. Original figure © 2018 Matiasz et al. [MWD18]; used here under CC BY 4.0 with a different font.	36
3.3	The research map of Figure 3.1 with its hypothetical edges removed. This modified research map, when compared with the one in Figure 3.1, illustrates how hypothetical edges help to structure research maps, thereby augmenting the interpretation of empirical results. © 2018 Matiasz et al. [MWD18]; licensed under CC BY 4.0.	37
3.4	The local map of ResearchMaps. The form on the left is used to input information. The citation on the top indicates the article whose research map is displayed. Highlighted in yellow are edges that reflect the article’s main ideas. Users can double-click on any edge to retrieve PubMed citations that are potentially relevant to the agent–target relation. © 2018 Matiasz et al. [MWD18]; licensed under CC BY 4.0.	40
3.5	The global map of ResearchMaps. The form on the left is used to query all the research maps in the application’s database. On the right is a panel that displays the research map returned in response to the query. © 2018 Matiasz et al. [MWD18]; licensed under CC BY 4.0.	41

3.6	Provenance of edges in the global map. Each edge in the global map can be clicked, revealing a table that lists every empirical result or hypothetical assertion recorded for that edge. Each entry in this table has a link to the local research map that contains the edge that was clicked. © 2018 Matiasz et al. [MWD18]; licensed under CC BY 4.0.	42
3.7	This is Alcino Silva’s personally curated research map of work in the field of memory allocation, as well as related work that either overlaps or connects to the work in memory allocation. To minimize the number of nodes, only the What property of each node is shown, so that nodes with different Where and When properties (but identical What properties) are collapsed into one. Nodes in orange appear only in research maps for articles on memory allocation. Nodes in red appear not only in research maps for articles on memory allocation but also in research maps of related work. © 2018 Matiasz et al. [MWD18]; licensed under CC BY 4.0.	43
6.1	A shorthand method for calculating the CEI for an edge in a research map. A table representing the model space of studies is instantiated with a pseudocount of one (a form of Laplace smoothing). The symbols along the left indicate the study classes involving an agent, A : positive intervention ($A \uparrow$), positive non-intervention ($A \emptyset^{\uparrow}$), negative non-intervention ($A \emptyset^{\downarrow}$), and negative intervention ($A \downarrow$). The symbols along the top indicate the results recorded in a target, B : increase ($B+$), no change ($B0$), and decrease ($B-$). This particular instantiation of the scoring table encodes four ($5 - 1$) positive interventions in which the target increased, one ($2 - 1$) positive non-intervention in which the target decreased, and one ($2 - 1$) negative non-intervention in which the target decreased. There are thus five studies suggesting an excitatory relation (green regions), and one study suggesting an inhibitory relation (red region). © 2018 Matiasz et al. [MWD18]; licensed under CC BY 4.0.	55
6.2	An example of an edge in a research map. This research map encodes three studies: two positive interventions (\uparrow) and one negative intervention (\downarrow). This edge is part of the research map for Han et al. [HKY07]. © 2018 Matiasz et al. [MWD18]; licensed under CC BY 4.0.	55

6.3	The growth of an edge’s CEI due only to consistency (left) and due to convergence (right). These plots show how the CEI of a research-map edge increases with each subsequent experiment (all with agreeing results), due to the principle of consistency (left) and due to the principle of convergence (right). The plot on the left represents repeated iterations of the same class of experiment (e.g., positive intervention) with consistent results. The plot on the right represents multiple iterations of experiments in which, at each iteration, one of the least-represented classes of experiments was performed, leading to consistent results. These two plots express an axiom of research maps: convergence carries greater epistemological weight than consistency. Original figure © 2018 Matiasz et al. [MWD18]; used here under CC BY 4.0 with modified axis labels.	57
8.1	A pattern of research-map edges that imply a lack of paths between the variables <i>A</i> and <i>D</i> . The Neo4j query for detecting a conflict with this pattern is given in Figure 8.2. Exactly this conflict was found in the original research map for a neuroscience article [DQQ15], as highlighted in Figure 8.4.	73
8.2	The Neo4j query used to find the conflicts that are highlighted in Figure 8.1, and which appeared in the original version of a research map (Figure 8.3).	74
8.3	Alcino Silva’s original research map for Do-Monte et al. [DQQ15], which contained conflicts. This research map was revised to resolve its conflicts, yielding the research map in Figure 8.4.	75
8.4	A revised version of Alcino Silva’s original research map for Do-Monte et al. [DQQ15] (Figure 8.3) that resolves conflicts present in the original version.	76
8.5	Alcino Silva’s research map for Giese et al. [GFF98].	77
8.6	Alcino Silva’s research map for Tsien et al. [THT96].	77
8.7	A research map showing a subset of the results in Figure 8.5 and Figure 8.6 merged into a single research map.	78

8.8	The “degrees of freedom” of an equivalence class. Each edge represents one of a causal graph’s <i>degrees of freedom</i> —i.e., one of the edge relations that can exist between two nodes in a causal graph. A dotted line denotes the relation in which the pair of nodes has no direct edge between them in the corresponding causal graph (e.g., $X \not\rightarrow Y$). Black edges are present in at least one causal graph in the equivalence class. Red edges are <i>not</i> present in the equivalence class, representing hypotheses that are inconsistent with the available evidence. The one red dotted edge—between “palphacalcin A” and “visual learning”—implies that every possible causal graph in the equivalence class has a direct edge between these nodes. This diagram demonstrates that even among the graphs that accommodate all the annotated constraints, many causal edges remain viable. Additional constraints—and thus additional experiments—are needed to eliminate edges from consideration. Note, however, that many edge relations have already been ruled out: the available evidence already precludes many edges (in red) from appearing in any of the consistent graphs. Such implications would be prohibitively difficult for a researcher to calculate by hand. © 2017 IEEE [MWW17b].	82
8.9	One of the thousands of optimal causal graphs derived from annotated results in literature. Each edge is a viable degree of freedom (Figure 8.8). © 2017 IEEE [MWW17b].	83
8.10	A comparison of three experiment-selection policies: (1) random, (2) Algorithm 2, and (3) Algorithm 3. This plot shows the results of the simulation given in Algorithm 4 for $N = 4$. The results show the experimental effort that is saved when each experiment is chosen based on the remaining degrees of freedom in the equivalence class.	84

LIST OF TABLES

2.1	The possible combinations of study classes and results, along with the relation that is implied in each case. The plus symbol (+) denotes an increase; the minus symbol (−) denotes a decrease; and a zero (0) denotes no change.	12
2.2	The number of possible DAGs over N variables, for $N = 1$ to 10.	21
4.1	The translation of research map annotations to (in)dependence relations for use in constraint-based causal discovery. This table includes research map annotations involving only one agent and one target.	46
5.1	The ASP encodings for the (in)dependence relations (i.e., Clingo’s input) in Figure 1.4. The variables X , Y , and Z are identified with the integer indices 1, 2, and 3, respectively. Weights are not assigned to these constraints; the parameter W is used as a placeholder.	50
5.2	The ASP encoding (i.e., Clingo’s output) for the causal graphs in Figure 1.4. The variables X , Y , and Z are identified with the integer indices 1, 2, and 3, respectively.	50
7.1	The possible patterns for degrees of freedom in an equivalence class over DAGs. Each pattern is associated with an interpretation given in plain language. In this table, “connected” implies “ <i>directly</i> connected”: note that all these patterns—including the first—still allow for an <i>indirect</i> path between X and Y via other nodes in the graph. A scientist can inspect these patterns to see which edge relations have been ruled out by the available evidence. This information can motivate the selection of additional experiments (Table 7.2).	66
7.2	The experiments that would be most informative with respect to a pair of variables, given their particular degree-of-freedom pattern in an equivalence class. These suggested experiments inform the experiment-selection methods given in Algorithms 2 and 3. In the third column, the variable that appears after each intervention symbol (\uparrow or \downarrow) is the agent that is intervened on.	70

8.1 The (in)dependence relations derived from the research map in Figure 8.7, along with their ASP encodings. For brevity, only the What property of each variable is listed. The following integer indices are used in the first two arguments of each ASP constraint: 1: p α -CaMKII T286; 2: LTD; 3: spatial learning; 4: LTP; 5: NMDAR; 6: visual learning; 7: NMDAR1. Because these constraints are satisfiable, each constraint was arbitrarily assigned a weight of 1. 79

ACKNOWLEDGMENTS

I am most grateful to my parents, Nestor John Matiasz and Maria Matiasz, who have sacrificed selflessly and inexhaustibly to ensure that I could take every educational opportunity that has emerged. This dissertation is dedicated to them.

My happiest day during graduate school was the day that my best friend became my wife: Gabrielle Green—of all those who have helped me—has been most patient in supporting me through the challenges of graduate school. Our time together has shown me that life’s most important lessons most certainly happen outside of the classroom. I am also grateful for the companionship of Gabrielle’s family, including Daniel Green, Lourdes Green, Julianne Green, Catherine Green, Stuart Green, Lisa Green, Murray Green, Dorothy Green, and Leandro Bravo.

More proximate causes of this dissertation are my two outstanding advisors, Alcino J. Silva and William Hsu, who have made my time in graduate school immensely rewarding. Their kindness, insight, and encouragement have been more than I could have asked for in advisors. I’ve enjoyed getting to know them first as my bosses, then as my colleagues, and now as my friends. Our one-on-one conversations in their offices are the highlights of my academic career and will likely remain my primary memories of my time at UCLA. I would also like to thank the rest of my dissertation committee—Alex Bui, Ricky Taira, and Denise Aberle—for their kindness, enthusiasm, and thoughtful feedback, which has improved this work considerably.

My honorary third advisor, Frederick Eberhardt, has been helpful, patient, and responsive. I’m thankful for his many suggestions, which have been instrumental in shaping the work presented in this dissertation. I would have been fortunate just to have found a world expert in causal discovery at nearby Caltech; a bonus turned out to be that he communicates with exceptional clarity and is tremendously generous with his time.

Much of the success in my research is due to my collaborator Justin Wood, whose creativity as a researcher and skill as a software developer have made my life immeasurably easier. In the body of this text, I’ve noted the ideas that would not exist in their current form without Justin’s valuable contributions.

I am grateful to everyone in the Medical Imaging Informatics Group and Silva Lab. They have been the ideal labmates: friendly, supportive, and encouraging.

I would like to acknowledge others who have helped and inspired me—either directly, through their friendship, or indirectly, through their work. These people, in roughly chronological order of my encountering them or their work, are Deborah Burns, William Burns, Thomas Burns, Adam Burns, Nicholas Burns, Charlotte Anderson, Martha Child, Jimi Hendrix, Kara (Murray) Vidal, Maria-Fatima Santos, Margaret Rosa, The Q, Bert Thurber, Frank Merrill, Dominic “Doc” Failla, Jeff Scanlon, Stephen LaBerge, B. Alan Wallace, Justin Miller, David Arond, Jason Yeager, Ron Lasser, Claire Rollor, Patrick Barber, Keith Tang, Laura Ogburn, George Nagel, Jael Jose, Jason Emmanuel, Katharine Brush, Noam Chomsky, Carla E. Brodley, Jonathan B. Freeman, Laura Henderson, Stephen N. Gomperts, Jean-luc Doumont, Albert J. Landini (who provides lucid commentary on much of my writing), Judea Pearl, Clark Glymour, Joseph M. Williams, Francis-Noël Thomas, Mark Turner, Matthew Butterick (whose fonts EQUITY, CONCOURSE, and *TriPLICATE* are used in this dissertation), and Bryan A. Garner.

Portions of this text have been published in my first-author papers, which are listed below under PUBLICATIONS. Some of the figure legends reference the Creative Commons Attribution License (CC BY 4.0), which can be obtained at <https://creativecommons.org/licenses/by/4.0/>. I intend to publish other portions of this text in papers that are currently in preparation or in submission. If you consider the contributions of my many co-authors, you see that it’s really quite silly for my name to appear alone at the beginning of this text.

One last point: If you read and even learn from this dissertation, I sincerely ask that you apply what you learn only if your intention is to reduce suffering. I acknowledge that this can be trickier than it seems.

Nicholas J. Matiasz
Los Angeles, CA
24 May 2018

VITA

2010	Bachelor of Science in Electrical Engineering <i>cum laude</i> , Tufts University
2010–2011	Research Assistant in Computer Science, Tufts University
2011–2012	Research Assistant in Psychology, Tufts University
2012	Master of Science in Electrical Engineering, Tufts University
2012–2013	Research Technician in Neurology, Massachusetts General Hospital
2013–2018	Graduate Student Researcher in Bioengineering, UCLA
2016	Master of Science in Bioengineering (Medical Informatics), UCLA
2018 (expected)	Doctor of Philosophy in Bioengineering (Medical Informatics), UCLA

PUBLICATIONS

Matiasz NJ, Hsu W, Silva AJ. ResearchMaps.org, a free web application to track causal information in biology. Poster session presented at: 13th Annual Molecular and Cellular Cognition Society Meeting, Society for Neuroscience Satellite Symposium; 2014 Nov 13–14; Washington, D.C.

Matiasz NJ, Silva AJ, Hsu W. Synthesizing clinical trials for evidence-based medicine: a representation of empirical and hypothetical causal relations. Poster session presented at: AMIA 2015 Joint Summits on Translational Science; 2015 Mar 23–27; San Francisco, CA.

Garcia-Gathright JI, **Matiasz NJ**, Garon EB, Aberle DR, Taira RK, Bui AAT. Toward patient-tailored summarization of lung cancer literature. In: Proceedings of the IEEE International Conference on Biomedical and Health Informatics (BHI); 2016 Feb 24–27; Las Vegas, NV. p. 449–52.

Matiasz NJ, Wood J, Hsu W, Silva AJ. ResearchMaps.org: A free web app for integrating and planning experiments. Poster session presented at: 15th Annual Molecular and Cellular Cognition Society Symposium; 2016 Nov 10–11; San Diego, CA.

Matiasz NJ, Chen W, Silva AJ, Hsu W. MedicineMaps: a tool for mapping and linking evidence from experimental and clinical trial literature. Poster session presented at: AMIA 2016 Annual Symposium. 40th Annual Symposium of the American Medical Informatics Association; 2016 Nov 12–16; Chicago, IL.

Matiasz NJ, Wood J, Wang W, Silva AJ, Hsu W. Computer-aided experiment planning toward causal discovery in neuroscience. *Frontiers in Neuroinformatics*. 2017 Feb 13;11:Article 12.

Matiasz NJ, Wood J, Wang W, Silva AJ, Hsu W. Translating literature into causal graphs: Toward automated experiment selection. In: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2017 Nov 13–16; Kansas City, MO. p. 573–576.

Garcia-Gathright JI, **Matiasz NJ**, Adame C, Sarma KV, Sauer L, Smedley NF, Spiegel ML, Strunck J, Garon EB, Taira RK, Aberle DR, Bui AAT (2017). Evaluating Casama: Contextualized semantic maps for summarization of lung cancer studies. *Computers in Biology and Medicine*. 2018 Jan 1;92:55–63.

Matiasz NJ*, Wood J*, Doshi P*, Speier W, Beckemeyer B, Wang W, Hsu W, Silva AJ. ResearchMaps.org for integrating and planning research. *PLOS One*. 2018 May 3;13:e0195271.

CHAPTER 1

Introduction

When scientists perform an experiment, they usually use a statistical method to show whether their experiment's result is significant. But when scientists select their next experiment, they rarely use a quantitative method to show whether their experiment's design is optimal. In much basic-science research, and in this dissertation, an experiment's *design* consists of two main choices: (1) the choice of which phenomena—out of all the potential phenomena in a system—will be involved in the experiment, and (2) the choice of which empirical strategy will be used—either a passive observation or an intervention where one or more of the phenomena are manipulated. For instance, given the available evidence, it may be more informative to intervene on variable X and observe the response of variable Y than it would be to observe whether variables Y and Z covary; in other situations—with different evidence available—the opposite may be true. There are still other situations where, given conflicting evidence, it may be most informative to repeat an experiment. Given the importance of such decisions, one can ask why scientists quantify the significance of an experiment's result, but not the potential significance of that experiment's design. Why is it that empirical results are usually assessed objectively, and empirical designs are often selected subjectively [SLB14, pp. 1, 42–43]?

This inconsistency in scientists' objectivity is a striking asymmetry in the scientific method. Despite the rigor, objectivity, and statistical validation that characterize scientific experiments, the cognitive process that occurs in a scientist's mind in between experiments—*experiment planning*—often happens informally and in an abstract way that cannot be shared with the scientific community. This subjectivity in experiment planning stands in stark contrast to the extreme measures scientists take to ensure the objectivity of the experiments themselves. Figure 1.1 depicts this peculiar bifurcation of the scientific method into objective and subjective components.

How can scientists select experiments more objectively? For those who want to identify a system's causal relations, this dissertation offers an answer: first, identify the causal explanations

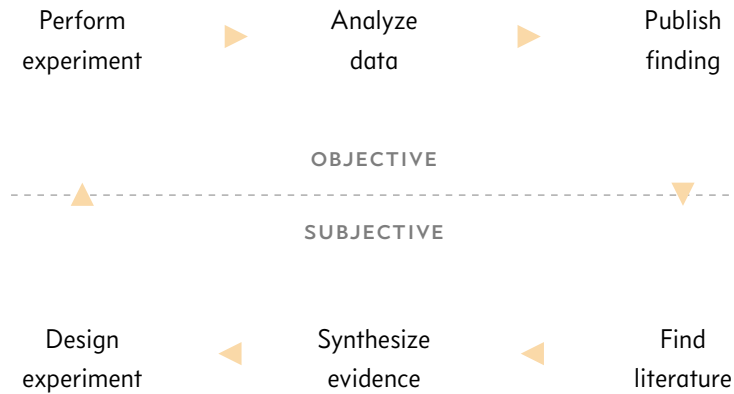


Figure 1.1: A depiction of the scientific process from the perspective of basic-science researchers who perform experiments (e.g., molecular biologists). See Clark and Kinoshita [CK07] and Russ et al. [RRH11] for alternative depictions.

that are consistent with the available evidence; next, select the experiment that could eliminate the most explanations from consideration. By taking maximum advantage of the available evidence to minimize their uncertainty, scientists can not only identify the smallest set of plausible explanations but also plan their next experiment more effectively [Fed72, p. 7].

Causal explanations can be expressed formally with the mathematical device of a *causal graph*, a directed graph in which the notation $X \rightarrow Y$ denotes that variable X in some way controls the behavior of variable Y [SGS00, Pea09]. Causal graphs visualize such relations concisely and intuitively, much like the diagrams of signaling pathways that pervade the biological research literature. But unlike most pathway diagrams, causal graphs also have precise and predictive mathematical properties, making them a more suitable representation for conveying not just empirical results but the *inferences* that one can make by considering combinations of empirical results.

My hypothesis is that scientists can quantify the value of potential experiments—and thus design experiments more objectively, with an analytic basis—by representing the implications of empirical results with causal graphs. This dissertation more directly addresses two supporting sub-hypotheses:

1. *The empirical results reported in research articles can be translated into constraints on the structure of a causal graph.* Although it is not yet common for many scientists to report their findings with causal graphs, research articles do commonly report statistical information, including statistical dependence and independence relations between phenomena. This dissertation shows how this statistical information can drive causal-discovery procedures that identify consistent causal explanations. The result is a literature-based, meta-analytic approach to causal discovery that can incorporate scientists' background knowledge and support experiment planning.
2. *Experimental design will be made more objective and communicable to the research community if each potential experiment is selected on the basis of its ability to reduce the underdetermination of a system's true causal graph.* To the extent that an experiment is designed to identify a system's causal relations, some experiments yield more information than others—depending on what is already known, and what is assumed about the system. An experiment's value can be made explicit and quantitative by using causal graphs to represent the causal explanations that are consistent with—and those that are ruled out by—the experiment's result (assuming, of course, that the result is correct). As scientists perform experiments, refuting specific explanations and homing in on the truth, causal graphs give a quantitative framework for minimizing the number (or cost) of experiments needed to identify the system's true causal graph.

1.1 Motivation: The imprecision of biological pathway diagrams

Causality is a primary concern in science, and particularly in medicine, where cause and effect determine matters of life and death. Although some researchers have proposed that we dispense with the concept of causality—and speak only of correlations—such movements seem to have lost favor: causality is now discussed in literature across many scientific disciplines [Her18].

We may not soon reach philosophical consensus on causality—much about the topic is still debated [Woo05, Reu13]—but we are certainly in an unprecedented time regarding causality's relationship to mathematics [Pea17]. The work of Judea Pearl, Clark Glymour, and others has allowed us to express causality using probability and statistics, yielding new technical applications and research programs [Pea09, SGS00]. These developments have led Dr. Gary King to assert, “More

has been learned about causal inference in the last few decades than the sum total of everything that had been learned about it in all prior recorded history” [MW15]. Scientists would do well to use these methods, but practical barriers remain to applying them in the laboratory. This dissertation presents practical causal-reasoning tools, tailored to the needs of scientists who want to reason more rigorously not just with data, but also with qualitative information from the research literature.

Causal graphs are a particularly suitable formalism for reasoning about biology because biologists tend to think in graphical terms. This preference for graphical image schemas is demonstrated by the ubiquity of pathway diagrams in the biological literature. In a pathway diagram, each node signifies a biological phenomenon, and each edge between nodes signifies a relation between phenomena. The result is a schematic summary of empirical results and their interpretations; Figure 1.2 is an example. These diagrams are useful in that they can concisely present complicated networks of interactions; as such, biology has more graphical information in its literature than do most other fields [LHM09]. It has even been suggested that the graphical depiction of a directed path from a source to a target (e.g., $S \rightarrow T$) is the most common image schema used to structure the expression of ideas [TT11, p. 64].

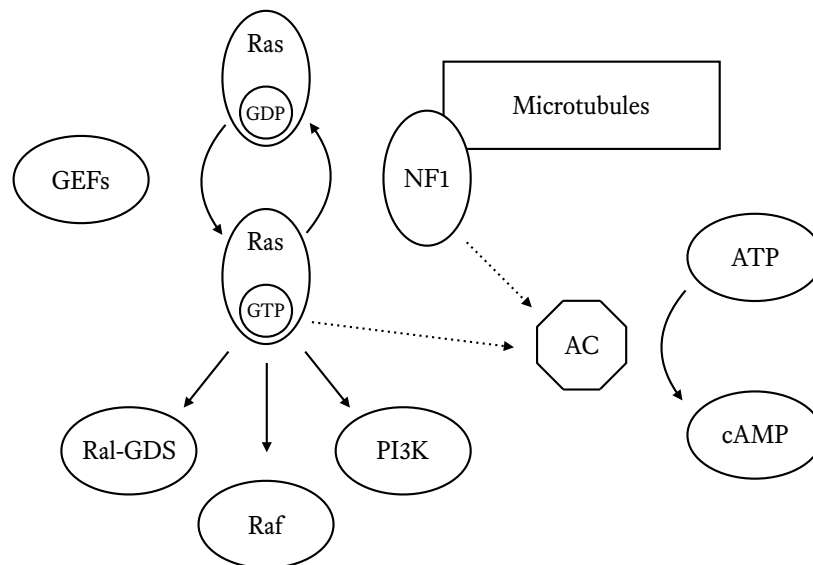


Figure 1.2: An example of a pathway diagram that has been adapted from a research article [CS03]. This diagram illustrates biological mechanisms, but because the meaning of each edge is not precisely defined, this diagram cannot necessarily be used to reason causally about the system.

However, biological pathway diagrams lack the kinds of standardized semantics and mathematical descriptions that have been developed for causal graphs. There is not one universal standard for expressing pathway diagrams, so a given diagram does not always give rise to one unambiguous interpretation [LHM09]. This problem is compounded when one attempts to synthesize pathway diagrams from multiple articles.

Multiple pathway diagrams cannot be synthesized by simply constructing the union of the diagrams' nodes and edges. As an example, consider Figure 1.3. The first pathway conveys that a change in X preceded a change in Z , with reason to believe that X in some way affected Z . The second pathway—say, from a separate article—presents a more nuanced picture: a change in X preceded a change in Y , and the change in Y preceded a change in Z . A biologist who encounters these diagrams in the literature may want to combine them—both to reduce the graphical information that they need to consider and to see what these diagrams imply when considered together. The third pathway in Figure 1.3 is a hybrid diagram that consists of the union of the first two diagrams. Note that because of the $X \rightarrow Z$ edge, it appears as though X can affect Z independently of Y , even if, for instance, Y 's activity is experimentally blocked. But this interpretation does not necessarily follow from the empirical evidence that led to the first and second diagrams. For example, it is possible that Y was unknown and thus unmeasured in the first study. Even if, in reality, Y mediated this $X \rightarrow Z$ interaction, it was not part of the explanation derived from the empirical evidence. So while this hybrid diagram may be valid, it is not the *only* diagram that accounts for the evidence: the second pathway in Figure 1.3 is another valid option. To know which diagram is correct, we would need to know more about the studies that led to these diagrams (if such information is even available), or we would need to perform additional studies. For instance, we could prevent Y from changing and see if intervening on X still causes a change in Z .

A key point is that empirical evidence can be perfectly consistent with multiple explanations, and thus with multiple pathway diagrams [Fry90, VP91, SGS00]. With every variable that we add to the system, the number of possible pathways grows super-exponentially. Therefore, the bookkeeping required to identify every consistent explanation becomes increasingly complex—far beyond what an individual scientist can be expected to consider manually. The development of causal graphical models in the last few decades has enabled algorithmic solutions to this problem.

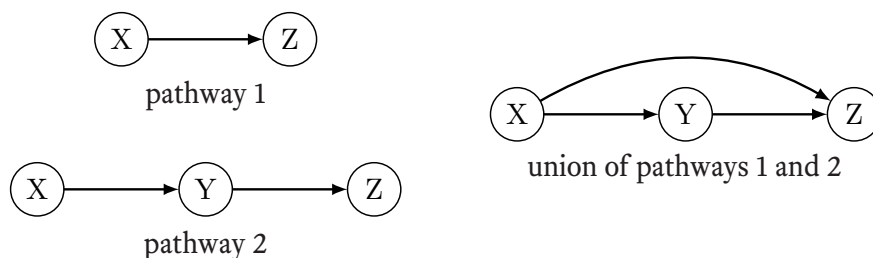


Figure 1.3: The union of pathway diagrams is not necessarily consistent with the evidence that is encoded in—or with the inferences that are allowed by—the individual pathway diagrams. Note that the $X \rightarrow Z$ edge implies that X can affect Z independently of Y , even though the evidence that gave rise to the original pathways does not guarantee that this interpretation is correct.

If pathway diagrams cannot be synthesized by taking their union, how then should they be combined? The approach presented here is to

1. identify the empirical evidence used to construct pathway diagrams;
2. translate this evidence into formal constraints on causal structure; and
3. input these constraints to a causal discovery algorithm to identify consistent causal graphs.

The mathematical theory that underlies causal graphs allows multiple causal graphs to be fused while ensuring that the hybrid graph is logically consistent with its individual components. Rather than analyzing a combination of datasets, this analysis works at the level of *structural information*, as described by Danks and Plis:

In general, we contend that evidence amalgamation is sometimes best addressed by thinking about the underlying structures that generated the evidence, as we can thereby sidestep some of the standard problems of evidence amalgamation. For example, one challenge in evidence amalgamation is the possibility of different background conditions in different experiments. As a practical example, suppose that X does not cause Y in any individual, but that the base rates of X and Y are both higher in population S_1 than in S_2 . If we simply merge data from S_1 and S_2 , then we will find an association between X and Y , which suggests a causal connection of some sort. If we instead merge the causal structures inferred from each dataset, then we will correctly learn that there is

no causal connection between them (since we will learn “no connection” from each dataset). [DP17]

1.2 Aims: Practical causal-reasoning tools for scientists

Before scientists reason causally about a body of evidence, they need to decide which evidence to trust in the first place. To help scientists integrate and quantify evidence, this dissertation presents the *research map* representation for empirical results and hypothetical assertions. This representation quantifies the methodological diversity on which scientific claims are based, taking into account not only the *ontological* information (i.e., *what* happened in a study) but also the *methodological* information (i.e., *how* this information came to be known). By quantifying both the consistency of individual lines of evidence and also the convergence (or triangulation) of multiple lines of evidence, this model of scientific consensus building addresses important gaps in current meta-analytic methods. This method is thus offered as a strategy for dealing with problems of p -value interpretability and the related “replication crisis” that is now commonly discussed in academic journals and popular media [GSR16, ASS18]. We also demonstrate research maps’ use as an annotation schema for capturing information pertaining to causal structure, which is used as input for constraint-based causal discovery.

There are a variety of causal discovery algorithms that operate on primary data; however, less explored is the problem of building causal models with only qualitative information from scientific communication, such as research articles. This is an important problem in that much of the evidence that a scientist encounters is qualitative: research articles and scientific presentations, for instance, are often unaccompanied by primary data but nonetheless convey important information that should inform experiment planning. This dissertation uses recent advances in constraint-based causal discovery to develop a pipeline that allows for causal discovery and experiment selection in the absence of primary data, using evidence from free-text research articles instead. An advantage of this approach is that experts’ domain knowledge can readily be incorporated into the pipeline, thus constraining the model space in ways that primary data often cannot. And when data is available, it can also be processed by the pipeline, helping to further determine causal structures. This pipeline was evaluated for its ability to yield causal inferences, which can be derived either from

an individual research article or from the synthesis of multiple articles. It was also evaluated for its ability to provide experiment-selection heuristics based on graphical representations of causality. Below, we discuss how practical limitations in scientific research come into conflict with notions of optimality in experiment selection. We consider experimental designs under constraints that are common in biology, particularly molecular biology: experiments that involve only two variables, in which neither or one of the variables can be intervened on (i.e., experimentally manipulated). For simplicity we assume causal sufficiency and acyclicity (§ 2.2); however, the causal discovery algorithm that we use can accommodate latent variables and cycles. Under these assumptions, we present experiment-selection heuristics that can make formal concepts of causality more practical for scientists who would like to apply them in their work. With simulations, we show how these heuristics can be used to increase the efficiency with which scientists obtain causal knowledge.

1.3 Overview: Planning experiments with causal graphs

Figure 1.4 is a system diagram of meta-analytic causal discovery and experiment selection, which comprises the following steps. First, scientific communication (e.g., literature) is annotated to produce a schematic representation—for our purposes, a *research map*—of empirical evidence. Second, this schematic representation is translated into a set of statistical relations, expressed as logical propositions. Third, these propositions, interpreted as constraints on causal structure, are input to a constraint-based causal discovery algorithm that identifies all the causal graphs consistent with the constraints. (In cases with conflicting constraints, the algorithm identifies the causal graphs that are maximally consistent.) Fourth, this set of candidate models is analyzed using the *degrees of freedom* of the candidate models (§ 6.2) to identify which next experiments could be most informative. Each component of this pipeline is described in a subsequent chapter.

This dissertation proceeds as follows: Chapter 2 reviews the literature on research maps, causal graphs, causal discovery, and experiment selection; gaps in this literature are discussed in light of this dissertation’s contributions—specifically regarding meta-analytic causal discovery and a calculus of evidence for combining scientific results. Chapter 3 presents ResearchMaps, a web application that entails a major component of this pipeline, and which provided data for the system’s analysis. Chapters 4–7 present each component of the meta-analytic technique in greater detail: I

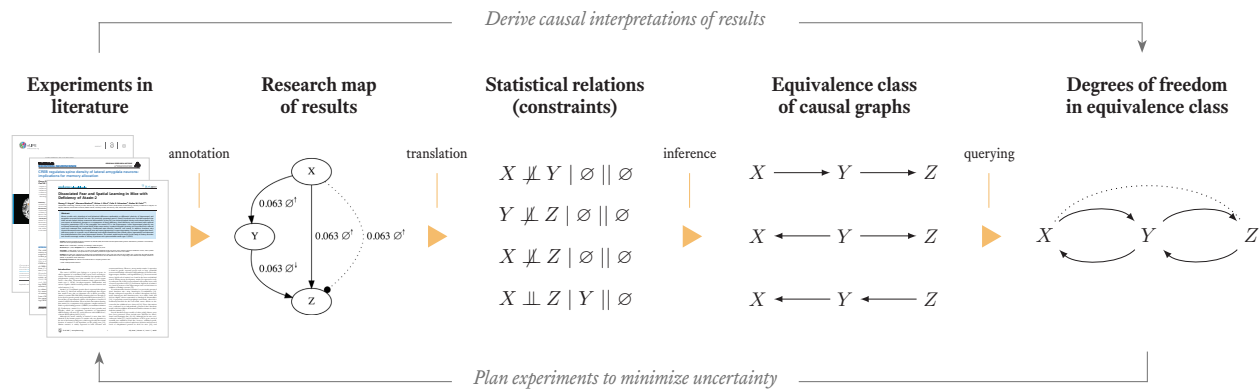


Figure 1.4: This system diagram gives an overview of a meta-analytic approach to causal discovery and experiment selection. (In the research map, the studies involving X and Z are shown on separate edges to highlight their correspondence to the third and fourth statistical relations.)

describe how to collect constraints on causal structure (Chapter 4), identify causal structures that are consistent with the constraints (Chapter 5), quantify evidence and causal underdetermination (Chapter 6), and select potential experiments by how informative they could be (Chapter 7). Chapter 8 presents evaluations and use cases of the system, designed to demonstrate the pipeline’s practical utility for biologists. Chapter 9 concludes with comments on my contributions, my hypotheses, the results’ generalizability, and the method’s range of applicability, as well as suggestions for future work.

CHAPTER 2

Literature review

2.1 Research maps

A *research map* is a graphical representation of empirical results and hypothetical assertions [LS13, SLB14, SM15, MWW17a, MWD18]. Figure 2.1 is a research map that represents a neuroscience article [CFK02]. As a graphical representation, a research map includes nodes and directed edges. Each node represents the identity and properties of a biological phenomenon, such as the protein CREB. A node can be identified with a simple text label, such as “CREB.” More formally, a node can be identified using the unique identifier (UID) for a concept within an ontology, such as the Gene Ontology (GO) [ABB00], the Unified Medical Language System (UMLS) [Bod04], or the Systematized Nomenclature of Medicine—Clinical Terms (SNOMED-CT) [Don06]. In the ResearchMaps web application (§ 3.1), each node is represented using the three properties of *what*, *where*, and *when*.

Each directed edge represents a relation between phenomena, such as the excitatory relation between CREB and spatial learning (CREB \rightarrow spatial learning). As a convention from biology, the node at the tail of the edge is called an *agent*; the node at the head is called a *target*. The agent for one edge can be the target for another. This agent–target image schema (*agent* \rightarrow *target*) is used to represent both empirical results and hypothetical assertions. When it represents empirical results, the schema conveys the result of a study that was actually performed. When it represents a hypothetical assertion, the schema conveys what a study’s result is hypothesized to be, should the study be performed. Hypothetical edges are drawn in a lighter color to distinguish them from empirical edges.

2.1.1 Categorizing evidence

Each instance of the agent–target image schema represents two kinds of information: *methodological* and *ontological*. The methodological information describes the empirical strategy used to study the agent and target—it conveys *how* facts about the phenomena are elicited from nature. For example, the agent may be experimentally manipulated; if so, this methodological aspect of the experiment should inform how we interpret the experiment’s results. This methodological information is categorized according to a taxonomy of empirical methods, which includes four classes:

- positive intervention (\uparrow)
- positive non-intervention (\emptyset^\uparrow)
- negative non-intervention (\emptyset^\downarrow)
- negative intervention (\downarrow)

In an intervention, an agent is experimentally manipulated, causing its quantity or probability to change; the target’s activity is measured to record whether it also changes, purportedly in response to the agent’s change. In a non-intervention—an observation—both the agent and target are passively observed, without intervention; the changes (or lack thereof) in both phenomena are recorded. In all four classes, “positive” and “negative” denote the direction of the change in the quantity or probability of the agent. The target’s quantity or probability may not change in a study, but the research-map schema requires the agent’s quantity or probability to have changed; otherwise, there could be no direct evidence of the agent’s effect on the target.¹

Studies in these four classes yield empirical results, which are represented in the ontological component of a research map. Whereas the methodological information describes *how* the results were obtained, the ontological information conveys *what* the study showed. For example, there could be an experiment where phenomenon *A* first increases, leading to an increase in phenomenon *B*. This result would imply a relation between the two phenomena.

¹ It is currently assumed that every edge in a research map is directional; therefore, the *agent* \rightarrow *target* schema is used even in non-intervention studies where the direction of causality cannot be posited on the basis of a correlation alone. This convention thus assumes that the researcher has background knowledge that is sufficient to posit a causal direction on the basis of this observational (non-intervention) result.

Relations between phenomena—the edges between nodes in a research map—fall into three categories:

- excitation (\rightarrow)
- inhibition (\dashv)
- no-connection ($\cdots\bullet$)

These relations usually imply the notions of positive correlation, negative correlation, and independence, but they are framed with vocabulary that is common among biologists. In a given study, the combination of the agent’s change and target’s change imply a specific relation. Table 2.1 presents the possible combinations of study classes and results, along with the relation that is implied in each case.

Study class	Change in agent	Change in target	Implied relation
positive intervention	+	+	excitation
positive intervention	+	0	no-connection
positive intervention	+	–	inhibition
positive non-intervention	+	+	excitation
positive non-intervention	+	0	no-connection
positive non-intervention	+	–	inhibition
negative non-intervention	–	+	inhibition
negative non-intervention	–	0	no-connection
negative non-intervention	–	–	excitation
negative intervention	–	+	inhibition
negative intervention	–	0	no-connection
negative intervention	–	–	excitation

Table 2.1: The possible combinations of study classes and results, along with the relation that is implied in each case. The plus symbol ($+$) denotes an increase; the minus symbol ($-$) denotes a decrease; and a zero (0) denotes no change.

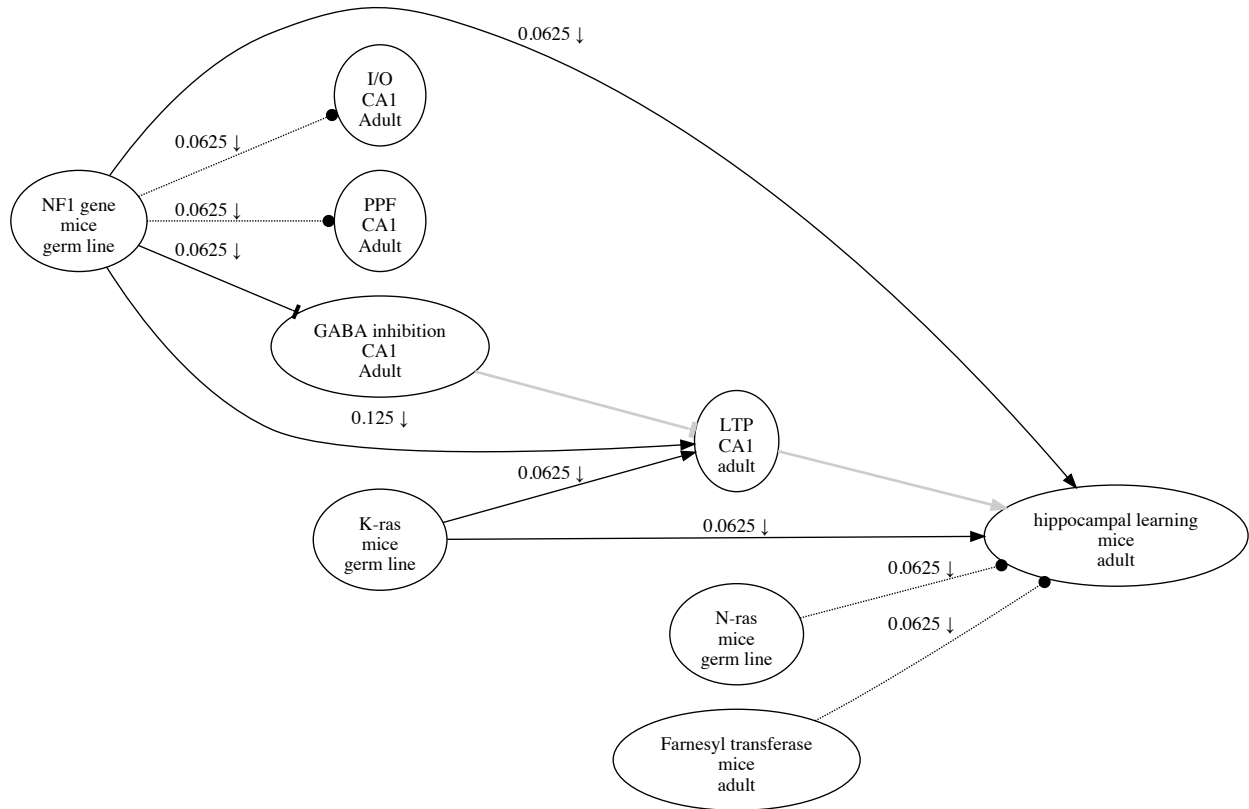


Figure 2.1: This research map represents the empirical results and hypothetical assertions reported in a neuroscience article [CFK02]. All three types of relations are shown: for instance, an excitatory edge from K-ras to LTP, an inhibitory edge from NF1 to GABA inhibition, and a no-connection edge from N-ras to hippocampal learning. The symbol on the edge from NF1 to hippocampal learning (\downarrow) indicates that at least one negative intervention was performed to test the relation between these two phenomena. The edges in gray—from GABA inhibition to LTP, and from LTP to hippocampal learning—are hypothetical edges: putative causal assertions that lack empirical evidence. Hypothetical edges are useful for incorporating assumptions or background knowledge about a causal system; they give the research map additional structure to facilitate the interpretation of empirical results. © 2017 Matiasz et al. [MWW17a]; licensed under CC BY 4.0.

2.1.2 Quantifying evidence

The research-map framework includes a method for quantifying evidence: each empirical edge is assigned a *cumulative evidence index* (CEI) on the interval (0,1) that conveys its evidence’s support

for a specific relation between the agent and target. The CEI is based on the idea that ontological information should be evaluated with respect to the methodological details of how it was obtained. For instance, a correlation between A and B may provide valuable evidence, but this evidence should be evaluated in light of whether the correlation manifested while A was experimentally manipulated in an intervention. There are at least two reasons for this. First, interventions and observations each have their own limitations regarding the information that they can provide. Observations can identify correlations in a system, but in most cases they cannot determine the direction of a causal relation—hence the popular refrain “correlation does not equal causation.” Interventions are useful for determining the direction of causality, but because they manipulate parts of the system and thus perturb it from its “natural” state, they are unable to identify some correlations—namely, those arising from causal paths that lead to the manipulated variable(s) [EGS06, Ebe07]. The second reason for considering ontological facts in light of their methodological context is that empirical methods are fallible: there is always the possibility for a study to yield a result that is a mere artifact. Scientists thus test hypotheses using a variety of methods to mitigate the risk of spurious results [SLB14].

The CEI is designed to express the epistemic principles of evidential *convergence* and *consistency*. By gauging the extent to which evidence is convergent and consistent, this scoring method helps to distinguish hypotheses with strong support from those with weak support. The principles of convergence and consistency are thus used for instantiating and scoring empirical edges in research maps.

Convergence analysis assesses whether the outcomes of the different kinds of studies (positive and negative interventions, and positive and negative non-interventions) are consistent with each other—i.e., whether they support a single relation type (either excitatory, inhibitory or no-connection). Suppose we find that optogenetically inhibiting cell type A is associated with a deficit in spatial learning. Suppose also that enhancing the activity of cell type A enhances the same form of learning. If we also found that cell type A is activated during spatial learning, and that this cell type is inactive when the animal is not learning, then our combined results would make a compelling argument that the activation of cell type A is causally connected to spatial learning. In a research map, this convergence between these four study classes would yield a relatively high CEI for the

excitatory edge between cell type A and spatial learning. On the other hand, contradictions among the data would lower the CEI of the edge. Convergence thus encompasses the notions that multiple lines of evidence are preferable to one, and that different study classes make unique contributions to testing the reliability of a hypothesized relation between two phenomena.

In addition to gauging the convergence of experimental results across multiple study classes, it is also important to gauge the consistency of empirical results within each study class. For this purpose, consistency analysis assesses whether experimental results are reproducible. For example, we might ask whether different kinds of positive interventions on the activity of cell type A (e.g., chemogenetic and optogenetic) always result in an enhancement of spatial learning. This question can refer to multiple iterations of the exact same experiment, or to a set of experiments that are similar in principle—e.g., two positive interventions of receptor A , one chemogenetic and the other optogenetic, which test two different forms of spatial learning.

In initial versions of the research-map framework, the CEI was calculated with a heuristic designed to express the principles of convergence and consistency [SM15]. This initial method worked as follows. Within each of the four study classes, the first study receives a score of 0.125. Each subsequent study in the same class receives a progressively smaller score according to a geometric progression, with an initial value of 0.125 and a common ratio of 0.5. For each study class, this geometric progression asymptotically approaches 0.25, such that the scores from the four study classes together can sum to a value in the interval (0,1). For example, the first positive intervention for a given agent–target pair receives a score of $0.25(1 - 0.5^1) = 0.125$. The second positive intervention receives $0.125/2 = 0.0625$, for a total score of $0.25(1 - 0.5^2) = 0.125 + 0.0625 = 0.1875$. This geometric progression expresses the principle of consistency, the idea that multiple replications of a study provide stronger evidence than just one instance of that study alone. But each replication contributes less than its predecessor because the results of successful replications are progressively less surprising. This progression of scores (0.125, 0.0625, 0.03125, . . .) is used independently for each series of studies within each study class. Treating each study class separately is an expression of convergence, the idea that multiple forms of evidence are always preferable to just one. Intuitively, each study class provides its own “perspective” on the hypothesis under consideration, helping to determine which of the possible relations has the dominant evidence. When the results of studies

conflict—e.g., some suggest an excitatory relation while others suggest an inhibitory relation—the total score of the edge is lowered by computing a normalized ratio that compares the dominant evidence’s score to the total score of all evidence. This method, whose derivation is given in Figure 2.2, has been replaced with a new one that expresses the same epistemic principles from a formal Bayesian perspective (§ 6.1).

	$B+$	$B0$	$B-$
$A \uparrow$	\mathcal{E}_{\uparrow}	\mathcal{N}_{\uparrow}	\mathcal{I}_{\uparrow}
$A\emptyset\uparrow$	$\mathcal{E}_{\emptyset\uparrow}$	$\mathcal{N}_{\emptyset\uparrow}$	$\mathcal{I}_{\emptyset\uparrow}$
$A\emptyset\downarrow$	$\mathcal{I}_{\emptyset\downarrow}$	$\mathcal{N}_{\emptyset\downarrow}$	$\mathcal{E}_{\emptyset\downarrow}$
$A \downarrow$	\mathcal{I}_{\downarrow}	\mathcal{N}_{\downarrow}	\mathcal{E}_{\downarrow}

$$\mathcal{E} = (1/4)(3 - 0.5^{\mathcal{E}_{\uparrow}} - 0.5^{\mathcal{E}_{\emptyset\uparrow} + \mathcal{E}_{\emptyset\downarrow}} - 0.5^{\mathcal{E}_{\downarrow}})$$

$$\mathcal{N} = (1/4)(3 - 0.5^{\mathcal{N}_{\uparrow}} - 0.5^{\mathcal{N}_{\emptyset\uparrow} + \mathcal{N}_{\emptyset\downarrow}} - 0.5^{\mathcal{N}_{\downarrow}})$$

$$\mathcal{I} = (1/4)(3 - 0.5^{\mathcal{I}_{\uparrow}} - 0.5^{\mathcal{I}_{\emptyset\uparrow} + \mathcal{I}_{\emptyset\downarrow}} - 0.5^{\mathcal{I}_{\downarrow}})$$

$$\mathbf{Score} = \frac{(\text{Max}(\mathcal{E}, \mathcal{N}, \mathcal{I}))^2}{\mathcal{E} + \mathcal{N} + \mathcal{I}}$$

Figure 2.2: The heuristic approach for calculating an evidence score, which was used in early versions of research maps. Each cell in the table starts with a value of zero, and each empirical result increments the appropriate cell by one. The function $\text{Max}(\mathcal{E}, \mathcal{N}, \mathcal{I})$ returns the maximum value from the set $\{\mathcal{E}, \mathcal{N}, \mathcal{I}\}$. The edge’s relation is assigned according to this maximum value: either excitatory (\mathcal{E}), no-connection (\mathcal{N}), or inhibitory (\mathcal{I}). © 2018 Matiasz et al. [MWD18]; licensed under CC BY 4.0.

Although a research map concisely summarizes a set of empirical results, it does *not* necessarily give the true causal explanation of those results [SLB14, pp. 130–133]. For instance, consider two studies, one where positively intervening on A produced an increase in B , and another study where positively intervening on A produced an increase in C . Considered together, these two studies could be represented by the research-map edges $B \leftarrow A \rightarrow C$. This diagram implies that A affects B in a process that is independent of the process by which A affects C . However, it is possible that the true causal path runs $A \rightarrow B \rightarrow C$, and that B was simply unmeasured in the study involving A and C . This is another example that shows how a set of empirical results can be perfectly consistent with multiple explanations (§ 1.1). Therefore, empirical results are represented with research maps, but causal explanations are represented with a different representation known as causal graphs.

2.2 Causal graphs

This dissertation expresses causality using the framework of causal graphical models [SGS00, Pea09]. This framework includes the notion of *causal structure*, which is a system’s particular configuration of causal relations that exist between phenomena. This network of causal relations is modeled by a *causal graph*, a directed graph $G = (\mathbf{V}, \mathbf{E})$ where \mathbf{V} is the set of vertices in the graph (variables in the system), and $\mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$ is the set of directed edges between the vertices in \mathbf{V} (causal relations in the system).² Relative to the vertices in \mathbf{V} , a directed edge in the graph (e.g., $x_i \rightarrow x_j$) conveys that the variable $x_i \in \mathbf{V}$ at the tail has a direct causal effect on the variable $x_j \in \mathbf{V}$ at the head. The *parents* of a particular variable x_j consist of every variable that has a direct edge from itself to x_j ; these parent variables can be thought of as the “variables Nature must consult before deciding the value of $[x_j]$ ” [Pea09, p. 203].

Causal graphs are described using the following graph terminology (many of these definitions are reproduced verbatim or nearly verbatim from [SGS00]):

Directed path A *directed path* from vertex A to vertex B in a graph G is a sequence of vertices beginning with A and ending with B such that for every pair of vertices X and Y that are adjacent in the sequence and occurring in the sequence in that order, there is a directed edge $E_{X,Y} = X \rightarrow Y$ in G .

Directed graph A *directed graph* has only directed edges.

Descendant A *descendant* of a vertex A is any vertex B such that there is a directed path from A to B .

Causal chain If there is a sequence of variables in \mathbf{V} beginning with A and ending with B such that, for each pair of variables X and Y that are adjacent in the sequence in that order, X

² It is instructive to distinguish between a causal graph and a causal model: A causal graph encodes only a system’s causal structure, the configuration of directed edges among the system’s variables, where each edge qualitatively signifies a causal relation. In addition to this structural component, a fully specified causal model has a parameterization, a quantitative specification of the values that each variable takes in relation to others. For example, a causal Bayesian network consists of both a causal graph that gives its structure and a set of conditional probability tables that gives its parameterization [Dar09]. This dissertation addresses the task of learning a system’s causal structure, as expressed by a causal graph.

is a direct cause of Y relative to \mathbf{V} , then we say that there is a *causal chain* from A to B relative to \mathbf{V} .

Source In a directed path from A to B , the *source* of the path is vertex A .

Sink In a directed path from A to B , the *sink* of the path is vertex B .

Acyclic path A path that contains no vertex more than once is *acyclic*; otherwise it is *cyclic*.

Directed acyclic graph (DAG) A *DAG* is a directed graph whose paths are all acyclic.

Mediator With respect to a directed path from vertex A to vertex C , vertex B is a *mediator* if it is on the path but is neither the path's source nor its sink.

Common cause A variable X is a *common cause* of variables Y and Z if and only if there is a directed edge $E_{X,Y} = X \rightarrow Y$ relative to $\{X, Y, Z\}$ and a directed edge $E_{X,Z} = X \rightarrow Z$ relative to $\{X, Y, Z\}$.

Collider With respect to a path in a graph, a *collider* on the path is a vertex whose adjacent edges both point toward the vertex. Vertex B is a collider on the path $A \rightarrow B \leftarrow C$.

Latent variable A *latent variable* is a variable that is unmeasured but causally connected to one or more variables in a system of measured variables.

Confounder A *confounder* is a latent common cause of two variables.

Causal sufficiency A set of variables is *causally sufficient* if there are no confounders.

A causal graph over a set of variables can be associated with a probability distribution over that same set of variables. When this association exists, the causal graph encodes features of the probability distribution, and vice versa. This association can exist given two assumptions:

- **Causal Markov condition:** A DAG G and its corresponding probability distribution $P(\mathbf{V})$ satisfy the *causal Markov condition* if and only if for every $x_i \in \mathbf{V}$, x_i is independent of its nondescendants, given its parents. Under this assumption, Reichenbach's *common cause principle* states that if x_i and x_j are statistically correlated, we know that (1) x_i causes x_j ; (2) x_j

causes x_i ; or (3) there is a set of common causes (or common causal ancestors) of x_i and x_j [RR56, Reu13]. These three conditions all individually imply that a path exists between x_i and x_j in the causal graph. Thus, under the causal Markov assumption, a probabilistic dependence implies a causal connection, and a causal separation implies a probabilistic independence [Ebe13].

- **Causal faithfulness condition:** A probability distribution is said to be *faithful* to its corresponding directed graph G if all and only the independence relations exhibited by the distribution are reflected in the causal structure of G . The assumption of causal faithfulness is the converse of the causal Markov assumption: under the causal faithfulness assumption, a probabilistic independence implies a causal separation, and a causal connection implies a probabilistic dependence [Ebe07]. Any independence exhibited by data generated from a faithful causal graph (and only these independencies) will be reflected in the structure of the graph. Without this assumption, probabilistically independent phenomena could still have a causal connection between them [Ebe13].

When the causal Markov and causal faithfulness conditions hold for a DAG, a useful correspondence exists between a graphical criterion known as *d-separation* and features of the probability distribution associated with the DAG: any conditional dependence or independence relation³ implied by d-separation holds if and only if the probability distribution also encodes this (in)dependence [GVP90]. D-separation can thus be used to read conditional (in)dependence relations off a DAG. The definition of d-separation is given in [Pea09, pp. 16–7] for disjoint sets of variables X , Y , and Z :

A path p is said to be d-separated (or blocked) by a set of nodes Z if and only if

1. p contains a chain $i \rightarrow m \rightarrow j$ or a fork $i \leftarrow m \rightarrow j$ such that the middle node m is in Z , or
2. p contains an inverted fork (or collider) $i \rightarrow m \leftarrow j$ such that the middle node m is not in Z and such that no descendant of m is in Z .

³ Below, I refer to such relations with the shorthand “(in)dependence relation” or “(in)dependence.”

A set Z is said to d -separate X from Y if and only if Z blocks every path from a node in X to a node in Y .

Under the Markov and faithfulness assumptions, if two (sets of) variables are d -separated, it follows that the (sets of) variables are independent in the probability distribution that is associated with the DAG. To determine if two disjoint sets of variables are d -separated, we can also ask if they are not d -connected. Hyttinen et al. [HEJ14] give the following definition of d -connection: “A path in graph G is d -connecting with respect to a conditioning set C if every collider c on the path is in C and no other nodes on the path are in C .” If a path is not d -connected, it is d -separated. This method for identifying d -separation is equivalent to Pearl’s method above [Stu98, Kos02].

The model space for causal graphs is enormous. The number D of possible DAGs that exist for N variables grows super-exponentially and is given by the following recurrence relation [Rob73]:

$$D(N) = \sum_{k=1}^N (-1)^{k-1} \binom{N}{k} 2^{k(N-k)} D(N-k) \quad (2.1)$$

This model space is relatively small for small numbers of variables: for sets of one, two, and three variables, there are one, three, and 25 possible DAGs, respectively. But for ten variables, there are over 10^{18} possible DAGs. To highlight how quickly this model space grows, Table 2.2 lists the number of DAGs that exist for one to ten variables.

2.2.1 Markov equivalence classes

Even though they have different graphical structures, two or more causal graphs can encode the same (in)dependence relations, as given by the rules of d -separation. A set of causal graphs that all imply the same (in)dependencies is called a *Markov equivalence class* [SGS00]. Figure 2.3 gives an example of a (Markov) equivalence class consisting of three graphs. Although the graphs’ edges have different orientations, they all imply the same (in)dependence relations and are thus observationally Markov equivalent: given only the observed (in)dependencies, the graphs are indistinguishable. An equivalence class thus formally expresses how evidence can be consistent with differing explanations, as discussed in § 1.1 and § 2.1.2. In the case of an equivalence class, the evidence can be a set of (in)dependence relations, and the differing explanations are expressed as causal graphs, each with a different causal structure.

Number of variables	Number of DAGs
1	1
2	3
3	25
4	543
5	29,281
6	3,781,503
7	1,138,779,265
8	783,702,329,343
9	1,213,442,454,842,881
10	4,175,098,976,430,598,143

Table 2.2: The number of possible DAGs over N variables, for $N = 1$ to 10.

This dissertation uses the phrase “equivalence class” in two ways: (1) to refer to a Markov equivalence class, as traditionally defined [SGS00]; and (2) to refer to the set of causal graphs that remain consistent with a set of evidence. Note that pieces of evidence can come into conflict with each other, and these conflicts can be resolved in multiple ways. Depending on how the conflict is resolved—and which evidence is discarded to achieve this resolution—different sets of graphs will be considered consistent. In this case, “equivalence class” is used to mean “the set of causal graphs that remain consistent with the evidence that one is currently willing to consider.” Throughout this dissertation, this phrase will be the intended meaning unless otherwise specified.

2.3 Causal discovery

The goal of *causal discovery* is to identify a system’s causal structure (i.e., its causal graph) given information that is derived from the system, such as (in)dependence relations between the system’s variables [Ebe17]. Causal discovery methods fall into three broad categories: (1) constraint-based methods, (2) score-based (or Bayesian) methods, and (3) methods based on semi-parametric assumptions [MD18].

Markov equivalence class	(in)dependence relations
$X \longrightarrow Y \longrightarrow Z$	$X \not\perp\!\!\!\perp Y \mid \emptyset \parallel \emptyset$
$X \longleftarrow Y \longrightarrow Z$	$Y \not\perp\!\!\!\perp Z \mid \emptyset \parallel \emptyset$
$X \longleftarrow Y \longleftarrow Z$	$X \perp\!\!\!\perp Z \mid Y \parallel \emptyset$

Figure 2.3: The three causal graphs on the left form a Markov equivalence class. Although their edges have different orientations, these three causal graphs all imply the (in)dependence relations on the right, in accordance with the rules of d-separation.

Constraint-based methods make use of the correspondence between (in)dependence relations in data and graphical structures in the associated causal graph. Data collected from the system is analyzed to obtain the (in)dependence relations that exist between the system’s variables. These (in)dependence relations are interpreted as constraints on causal structure; the task is then to identify the causal graphs that are (maximally) consistent with these constraints. Examples of constraint-based methods include the PC algorithm and the Fast Causal Inference (FCI) algorithm [SGS00]. Some constraint-based methods use Boolean satisfiability (SAT) solvers [BHM09] to identify consistent causal graphs. This dissertation relies heavily on one of these SAT-based methods [HEJ14].

Score-based methods define a measure that quantifies the fit between data and causal graphs. These methods commonly use a Bayesian approach: the task is to identify the causal graph that maximizes the likelihood of the data given the causal graph. An example of a measure that is used is the Bayesian Information Criterion (BIC) [Sch78].

Methods in the third category of causal discovery use semi-parametric assumptions to identify causal graphs with more efficiency or specificity. These methods differ from those in the first two categories in that they do not rely on the assumption of faithfulness. As an example of a semi-parametric assumption, one can assume that a system is governed by linear functions with non-Gaussian noise and use Linear Non-Gaussian Model (LiNGaM) algorithms to identify the system’s causal graph using independent component analysis [SHH06, Shi14].

Causal discovery methods are increasingly being used in the biological sciences. They have been used to identify protein-signaling networks [SPP05], cell signal transduction from proteomics experiments [III16], transcriptional regulatory networks [CES07], causal effects of genetic variants [MCK10], associations between gene expression and disease [SLY05], genetic mutations that will cause predictable phenotypic changes [SMS12], single nucleotide polymorphisms (SNPs) that predict disease [ALA11], and causal effects of environmental factors on genetic diversity between populations [FPP18].

2.4 Experiment selection

*Experiment selection*⁴ refers to the strategies that researchers use to design their next experiment. These decisions are generally affected by many factors, including research funding, laboratory resources, and investigators' interests. With respect to the goal of understanding a system's causal relations, experiment-selection techniques seek to maximize causal knowledge with a minimum of experimental effort. These techniques ask: which next experiment or sequence of experiments would most fully and efficiently determine the causal relations that govern the system's variables? Experiment selection can be either *fixed* or *adaptive*. A fixed procedure selects one specific sequence of studies before any are performed. An adaptive procedure is permitted to update its planned sequence of studies in response to the results of previous studies in the sequence [Ebe07].

Researchers have approached experiment selection using a variety of techniques. Murphy [Mur01] and Tong and Koller [TK01] take a Bayesian approach to identifying the best experiment to perform next. Given a prior distribution over possible DAGs (without latent variables), they enumerate the possible experiments one could perform next, compute a posterior distribution over the graphs that could result, and identify the experiment that maximizes information gain. Although this method is a principled Bayesian approach, it is very computationally expensive and thus does not scale well.

Meganck et al. [MML05] and He and Geng [HG08] use decision-theoretic heuristics to identify optimal experiments. Considering DAGs without latent variables, these researchers construct

⁴ In this dissertation, I use “experiment selection,” “experiment planning,” and “experimental design” interchangeably. That which is being selected, planned, or designed are the parameters of an experiment described in Chapter 1.

utility functions and decision criteria based on the number of underdetermined edge orientations (e.g., in a Markov equivalence class) that could be determined by subsequent experiments. Note that Meganck et al. use the empirical distribution of edge orientations in an equivalence class to estimate the true probability of orientations for each edge; this method is related to the experiment-selection algorithms given in § 6.2.

Graph-theoretic approaches to experiment selection have also shown promising results. Researchers have found that the problem of optimal experiment selection can be formulated as graph-theoretic and combinatoric problems [HEH13]. Using such formalisms, Eberhardt et al. [Ebe05] and Hyttinen et al. [HEH13] derive bounds on the number of experiments sufficient and in the worst case necessary to identify a causal graph uniquely. In addition to DAGs without latent variables, these researchers consider causal structures with latent variables and those with feedback (cyclicity); they also give algorithms for constructing an optimal *sequence* of experiments under various constraints, such as limiting the maximum number of variables that can be intervened on simultaneously. Eberhardt [Ebe08] and Hauser and Bühlmann [HB12] consider acyclic graphs without latent common causes and give algorithms for efficiently selecting optimal intervention sets. Such methods thus translate the semantics of experiment planning to well established methods in the literature on graph theory and combinatorics.

2.5 Gaps in the literature

2.5.1 Meta-analytic methods that quantify evidential convergence

Scientists use various heuristics to evaluate evidence and develop confidence regarding the truth of hypotheses. But this confidence is always achieved with inference procedures, and with incomplete evidence. This fact should not be taboo. Consider the alternative: if a hypothesis could not be deemed true until it was studied *exhaustively*—however strictly that might be defined—progress in science would be brought to an almost stagnant pace. For instance, scientists cannot hope to test every possible relation under every possible experimental context or condition, using every possible subset of variables; the combinatorics involved make this an impossibly expensive and time-consuming strategy [Dan05]. This is why scientists need to rely on some sort of inference—and,

indeed, why they already are doing so.

What exactly are the rules for this inference? Natural candidates include statistical measures, one of the most common being the p -value. But the p -value's ubiquity in science is incongruent with the amount of debate over its utility and misuse, which is now documented regularly in the literature [Gel13, GP13a, GP13b, HCV15, GSR16, AKR17]. Even the U.S. Supreme Court has ruled on this issue, agreeing unanimously in 2011 that “statistical significance is neither necessary nor sufficient for determining the scientific or practical significance of a set of observations” (Matrixx Initiatives, Inc. et al. v. Siracusano et al. No. 091156. Argued January 10, 2011, Decided March 22, 2011) [GSR16]. The enormous range of views on this issue suggests that science would benefit significantly from additional theoretical clarity on the p -value's role in science.

Despite concerns over the use of p -values and other statistics, the field of meta-analysis has used such measures for decades to synthesize empirical findings quantitatively. Introduced in its modern form in the 1970s, meta-analysis usually serves one of two goals: The first is to evaluate evidence—from a relatively small group of studies—for whether an intervention is effective in addressing a problem, often in clinical settings. The second goal is to generalize empirical findings—from a relatively large group of studies—yielding a more complete perspective than any individual study can offer in isolation [GKN18]. Meta-analysis thus helps researchers to assess the consistency of evidence and stands as the most sophisticated method of evidence synthesis currently available.

There is growing consensus that meta-analysis should more explicitly analyze and quantify *triangulation*, the use of several different methods—each with its own empirical strategy and potential sources of bias—to obtain evidence for a specific hypothesis [SW00, LTD16, MS18]. This concept is related to the notions of *intervention complexity* [NGL13, LHC17], *methodological diversity* [Joh03, Zol10], and *evidential convergence* (§ 2.1.2 and § 6.1). Meta-analysis offers sophisticated methods for quantifying the *consistency* of evidence, including measures of heterogeneity for effect sizes. However, there has been relatively little development of methods for quantifying triangulation, even though this concept has long been described qualitatively and acknowledged for its importance [WCS66, Smi81]. A recent review of evidence-synthesis methods for health and social policy found only one approach [Org14] that “[extends] the domain of consistency to consider evidence from different study designs” and “looks at evidence from different methodological approaches to

inform the rating of the quality of a body of evidence” [MDR18].

It is possible that triangulation is following a historical trend that has played out for other scientific concepts: one in which qualitative intuitions about a scientific concept are gradually translated into increasingly quantitative models that formalize the original concept while also preserving the qualitative features that made it instructive in the first place. An example of this trend is the development of increasingly refined mathematical models for causality, an old concept that has been explored by philosophers for centuries. David Hume, for instance, provided an influential definition of “cause,” as well as qualitative descriptions of related concepts, such as the copy principle and the problem of induction [Hum03, Hum16]. Qualitative intuitions such as these began to be formalized by Sewall Wright when he introduced path diagrams as a way to express causal associations [Wri21, Wri23, Wri34]. Additional qualitative concepts such as Austin Bradford Hill’s criteria for causation [Hil65] and Reichenbach’s common cause principle [RR56] further explicated notions of causality. These ideas have now been formalized to a greater degree with the framework of causal graphical models [SGS00, Pea09], a quantitative representation that formalizes ideas previously described only qualitatively.

If it is true that, like causality, the concept of evidence is following a similar historical trend, it is then less surprising that there would be such debate about the replication crisis: the crisis hinges on evidence’s consistency, which meta-analysis explicitly quantifies. Because we now can quantify consistency, we can scrutinize it to a greater degree. But perhaps less of a crisis would be perceived if meta-analysis also quantified triangulation with the same precision. Even if scientists fail to replicate a single line of experimentation, the totality of evidence might still point to a consistent set of explanations [SBS18]. Regardless of the effect that triangulation will ultimately serve in the replication crisis, science seems primed to develop methods for quantifying it [MS18].

Part of the challenge in quantifying triangulation is articulating what the categories of evidence should be: what exactly distinguishes the lines of evidence that are meant to converge, allowing us to “triangulate” our understanding of a system? Randomized controlled trials (RCTs) are usually seen to be the gold standard for causal inference, whereas other, nonrandomized studies take lower positions in so-called *evidence hierarchies* [MDR18]. But RCTs are simply infeasible in many research domains, often due to ethical concerns; for instance, we cannot (and should not) force participants

to smoke cigarettes. And recent developments in causal modeling have cast doubt on whether RCTs can independently provide the gold standard for inferring causal relations [Ebe13, Pea18]. A variety of methods are therefore needed, whether the justification is theoretical or pragmatic.

Quantifying triangulation is part of a larger challenge of quantifying evidence. Like causality, the concept of evidence is central to human inquiry. But, as was true with causality until only recently, evidence is still not quantified formally in most scientific research—at least not with widespread consensus as to how this calculation should be defined [VC18, MS18]. There have been many attempts to develop methods for quantifying evidence, with motivations grounded in probability, statistics, and information theory [Goo60, Goo67, Vie06, Lee11, Vie11, VH11, VDH13, Eva15, VS15, Eva16, VS16]. Debate over the replication crisis makes clear that this issue has not been resolved, as we have yet to articulate objective definitions of evidence that allow a proposition to be verified or refuted conclusively—with the authority and objectivity, for instance, that is attributed to measurements of temperature [Vie06]. In this dissertation, I take the position that quantifying triangulation, or convergence (§ 2.1.2), will bring us closer to a more complete definition of evidence that expresses epistemic principles already used by scientists.

2.5.2 Causal discovery without primary data

Regardless of how it might be quantified, causal evidence comes from published studies whose results are often disseminated only as free text in research articles, often in the form of aggregate statistics. To build a model of a causal system, a scientist must integrate these results with each other and with background knowledge. This qualitative information must also be integrated with knowledge gleaned from the analysis of primary data, when available. Scientists would thus benefit from meta-analytic causal discovery methods that can accommodate all the various forms of evidence that they encounter [Dan05, MWW17b].

Much of the current literature on causal discovery gives methods to identify causal relations using primary data from empirical studies (§ 2.3). Efforts such as the Center for Causal Discovery [CBB15] have developed robust causal discovery algorithms that operate on large-scale datasets, and much is now understood about *data fusion*—combining and learning from multiple datasets that were collected under different empirical conditions [BP16]. But it is not obvious how to gen-

eralize these methods to meta-analytic techniques that can incorporate multiple forms of causal information, including qualitative knowledge from published literature. This dissertation shows how constraint-based causal discovery methods (§ 2.3) provide a good platform for integrating qualitative evidence in free text.

2.5.3 Interpretable experiment-selection strategies

It was just in the last few decades that causality was formalized mathematically [PM18]; it was even more recently that researchers have used causal graphs as an analytic basis for proposing efficient experiment-selection criteria (§ 2.4). Although much is now understood about experiment selection, including its relation to combinatorics [HEH13] and decision theory [MML05], work is needed to translate the available theory and algorithms into practical tools that fit into scientists' current workflows [Gly04, KRO09]. David Danks nicely summarizes the task at hand [Dan05]:

For the experiment choice problem, a simple naïve algorithm would first enumerate the possible sequences of experiments as well as the possible integration outcomes for each stage in each sequence. We could then apply the above inference rules to each extended experiment-outcome sequence to determine the stage at which we would settle on a unique integrated structure. If we then had some probability distribution over the experiment-outcome sequences, we could determine which experiment sequence has the earliest expected stage at which it settles on a unique model. Of course, this strategy is hopeless from a computational point of view, because it requires both the enumeration of a highly exponential number of sequences and a specification of the probability distribution over experiment-outcome sequences. We can avoid the computational explosion by using some heuristic strategy, but that strategy will not be guaranteed to find the optimal experiment sequence. *Unfortunately, we must—in this domain, as in many others—make a decision between asymptotic correctness and computational tractability, and the balancing point for that trade-off depends on the particular domain and scientists.* (emphasis added)

Experiment-selection methods often use simplifying assumptions that are understandable given the complexity of the analysis, yet difficult to translate into practice. For instance, to sim-

plify the analysis, methods will often assume that the true causal graph for the system is acyclic (e.g., [TK01]). Other methods assume that global bounds are set on the number of variables that can be either observed or intervened on in any one study (e.g., [May13]). And other methods require each potential experiment to be assigned a cost, allowing for an objective function that is to be minimized by the experiment-selection policy (e.g., [MLM06]).

In reality, constraints on empirical work in the laboratory are far less global, and much more heterogeneous across experiments. For example, a particular experimental design may allow for a simultaneous intervention on three variables and an observation of six; in another experiment with a different design, researchers may be able to observe only two variables simultaneously [May13]. This heterogeneity of constraints lessens the relevance of the known bounds on the number of experiments that are sufficient and in the worst case necessary to identify a system's causal graph [Ebe05, EGS06, Ebe07, HB12]. Additionally, it is often infeasible for a scientist to objectively assign costs to potential experiments: the relevant constraints in such decisions involve more than just the monetary expenses of the necessary lab equipment, and often include subjective criteria that are difficult to quantify. In general, research decisions are constrained by practical issues of funding, timing, resources, and personal motivations. Thus, scientists could benefit from efficient heuristic methods for selecting experiments that accommodate the complexity of scientific research while still yielding instructive recommendations.

In the above quotation, Danks highlights a choice between asymptotic correctness and computational tractability, but I submit that there is a third issue in experiment selection: the question of whether scientists can interpret and see the rationale for an algorithm's experiment suggestions. The theoretical justification for experiment-selection algorithms are of course well founded if they are based on a sound mathematical understanding of the relevant theory, such as combinatorics [HEH13]. But the justification for the algorithm's suggestions will also be expressed in these particular abstractions. These abstractions may be inaccessible to a scientist who has an advanced understanding of the system under consideration but who nonetheless lacks training in the requisite mathematics. This is a significant obstacle because in most cases scientists make the final decision regarding which experiment to perform next. As a result, experiment suggestions whose rationale can be expressed only at the level of unfamiliar mathematical abstractions may be less persuasive

than suggestions grounded in representations that domain experts already use to conceive of systems, such as graphical image schemas like pathway diagrams (§ 1.1). Indeed, the tasks of analyzing evidence and designing experiments cannot be fully decoupled, as it is usually the gaps in evidence that primarily motivate the design of a scientist’s next experiment. This dissertation offers interpretable metrics defined over graphical representations of empirical evidence and causal structures, as well as heuristic experiment-selection methods based on these metrics. The heuristics will not outperform state-of-the-art experiment-selection methods that approach or achieve theoretical limits on performance, but they are designed to give experiment suggestions whose rationale can be readily interpreted by a scientist.

2.6 Contributions of this dissertation

2.6.1 The cumulative evidence index (CEI)

The heuristic approach for scoring evidence that is described in § 2.1.2 has been formalized using Bayesian statistics, yielding the *cumulative evidence index* (CEI). This new method is presented in § 6.1 and has been implemented in the ResearchMaps web application, which is presented in Chapter 3. The CEI models scientific reasoning as a type of distributed Bayesian inference [Kra17], providing a nuanced analytic basis for characterizing how scientists build consensus in their fields. It also addresses the lack of convergence in meta-analysis (§ 2.5.1) by quantifying it explicitly, and by allowing scientists to express the greater importance it holds relative to evidential consistency. Analyzing this model led to the definition of evidential *divergence* (§ 6.1), another epistemic principle that complements evidential convergence and consistency [MWD18]. This model thus contributes to meta-research [IFD15] by translating qualitative principles of scientific reasoning into quantitative parameters of a mathematical model, which can be analyzed and thus made more efficient through theoretical work, simulations, and historical meta-analyses of the scientific record.

2.6.2 A literature-based technique for causal discovery

To augment what can be learned from data-driven causal discovery, this dissertation presents a meta-analytic approach that allows researchers to identify causal structures using published findings in the

literature, even if they are not accompanied by primary data. This method uses machine-readable representations of empirical results in free-text resources like PubMed, which often give only limited, aggregate statistics. The strategy is to first translate free-text descriptions of empirical results into formal constraints on causal structure. These causal-structure constraints are fed as input to a state-of-the-art, constraint-based causal discovery algorithm [HEJ14]. This algorithm can compute *every* causal graph that is consistent with the constraints. Features of this model space—the set of causal interpretations that remain viable—are visualized and quantified for further analysis to facilitate experiment selection.

By grounding literature synthesis in the formalism of causal graphs, this method offers a number of benefits that could improve the rigor with which scientists evaluate evidence and select experiments. First, by allowing scientists to express empirical findings as formal constraints on causal structure, there is a clear demarcation between facts that are demonstrated empirically and background assumptions that are used to simplify the analysis. Because a constraint-based method is used, background assumptions can also be expressed as formal constraints to facilitate the search over causal graphs. For example, a domain expert may specify that in any causal path, a specific subset of variables should always come before another subset [Ebe17]. Alternative background assumptions can readily be substituted—leaving the empirical constraints intact—to evaluate the effect that the assumptions have on the set of consistent causal explanations. A second benefit is that scientists can query the system to confirm whether a hypothesis that they propose is a logical extension of previous results. Once a set of constraints has been obtained, further constraints that represent hypothetical results can be tested for logical consistency. If a hypothetical result introduces a logical conflict, the scientist can be confident either that the current set of constraints includes an incorrect proposition or that the hypothesized result is incorrect (or both). This categorization of hypotheses has enormous implications for experiment planning but is usually impractical without such tools. Having examined this information, scientists can select their next experiment with a more precise understanding of what has already been discovered.

2.6.3 Interpretable heuristics for experiment selection

To provide experiment suggestions whose rationale can be readily interpreted by scientists, this dissertation offers interpretable metrics for evidence and uncertainty defined over the graphical representations of research maps and causal graphs. With research maps, we quantify evidence using the cumulative evidence index and frame experiment selection as the maximization of empirical evidence for a specific hypothesis about causal structure. With causal graphs, we quantify the underdetermination of causal structure using the degree-of-freedom metric; we frame experiment selection as the minimization of this causal underdetermination. Heuristic approaches to experiment selection are defined with respect to these metrics, whose intuition is readily expressed with a vocabulary that scientists recognize. In these ways, scientists can select their next experiment with the same abstractions that they use to synthesize past results.

CHAPTER 3

ResearchMaps: a web application for experiment planning

ResearchMaps is a web application for building and querying research maps [MWD18]. It is currently hosted at <http://www.researchmaps.org> and is free for users at colleges, universities, and non-profit research centers. The application consists of two main pages: the local map and the global map. The local map shows the research maps for specific articles; it is where the author of the research map can modify it (Figure 3.4). The global map can be used to query the entire ResearchMaps database for specific phenomena and connections between them (Figure 3.5).

3.1 Implementation of the research-map framework

In ResearchMaps, an agent or target is defined in three complementary ways: *what* the phenomenon is, *where* the phenomenon exists, and *when* the phenomenon acts. ResearchMaps stores this information as three properties for each node: (1) *what* describes a key identifier of the phenomenon involved (e.g., the name by which the gene, protein, cell, organ, behavior, etc. is known); (2) *where* describes the location of the *what* (e.g., the organ, species, etc.); and (3) *when* provides temporal information that is critical to the identity of the *what* (e.g., the time, age, phase, etc.). For example, if the protein neurofibromin is measured in multiple locations, a corresponding research map would include multiple nodes for neurofibromin with different *where* properties. This approach is instructive, as neurofibromin could have different biological characteristics in different cellular locations (e.g., excitatory neurons versus inhibitory neurons) or at different stages of development. ResearchMaps displays the *what*, *where*, and *when* properties on separate lines within each node.

The four study classes are represented by symbols above each empirical edge. As defined in § 2.1, positive interventions are represented by an upward arrow (\uparrow); negative interventions are represented by a downward arrow (\downarrow); positive non-interventions are represented by the empty set

symbol and a superscript upward arrow (\emptyset^\uparrow); and negative non-interventions are represented by the empty set symbol and a superscript downward arrow (\emptyset^\downarrow). Although we have not yet defined a formal representation for experiments involving more than two nodes, ResearchMaps can store intervention experiments involving two agents. At the time of this writing, such experiments comprise approximately fourteen percent of the experiments logged. The putative mechanisms underlying the results of these multi-intervention experiments can be visualized using hypothetical edges among the three entities involved (two agents and one target); the structure of these hypothetical edges is provided by the user.

ResearchMaps can store information about the statistical test used to establish each finding, as well as its associated p -value. Such information is of course valuable in evaluating studies; however, as the areas covered by research maps are diverse, and there are no standards as to which statistics are used and how to report them, p -values do not currently affect the CEI of each empirical edge, and they are optionally tracked by each user. See Figure 3.1 for an example of a research map.

ResearchMaps allows the user to input both empirical and hypothetical edges between any two phenomena (and, by extension, empirical and hypothetical nodes). As introduced in § 2.1, a hypothetical edge represents a putative connection with no direct empirical evidence. Hypothetical edges are often implied by empirical edges, and they are often key in interpreting and reporting the results of a research article. As hypothetical edges do not represent empirical evidence, they are assigned neither CEIs nor study symbols. To visually differentiate hypothetical edges, they are shown in a lighter color and without these annotations on their edges.

Beyond allowing users to track hypotheses, hypothetical edges can also help to structure research maps of empirical evidence. Just as hypotheses help to frame and organize the results of research articles, hypothetical edges help to structure and contextualize empirical edges in a research map. Consider the hypothesized pathway $A \rightarrow B \rightarrow C \rightarrow D$. A research map that represents the empirical edges $A \rightarrow C$, $A \rightarrow D$, and $B \rightarrow D$ would not explicitly reflect the putative $A \rightarrow B \rightarrow C \rightarrow D$ pathway because not all connections in this pathway are part of that map. By including the hypothetical edges $A \rightarrow B$, $B \rightarrow C$, and $C \rightarrow D$, the underlying hypothesis for the performed studies is immediately obvious (Figure 3.2). To further illustrate this point, Figure 3.3 displays the research map of Figure 3.1 without its hypothetical edges.

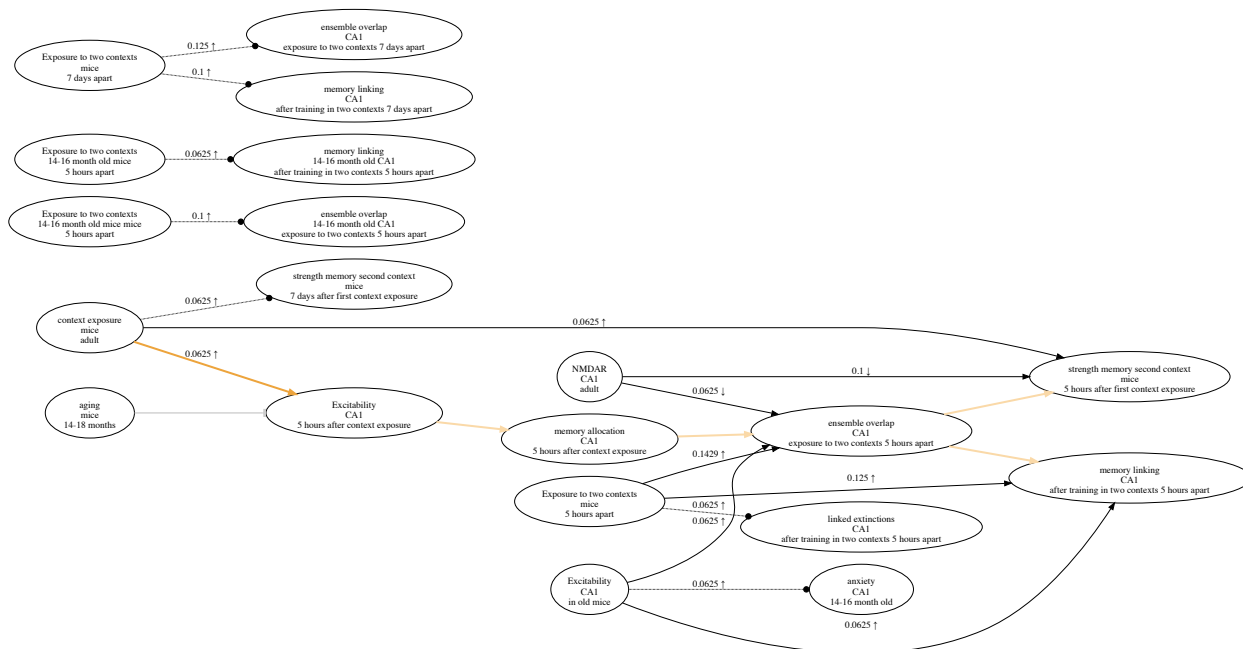


Figure 3.1: A research map of a published article [CAS16]. Each node in a research map has three properties: what (top), where (middle), and when (bottom). Nodes are connected by edges that represent exposure relations: excitatory (sharp arrowhead), inhibitory (blunt arrowhead), and no-connection (dotted line, circular arrowhead). Each empirical edge also has a CEI that reflects the amount of evidence represented, as well as symbols that reflect the study classes recorded for that edge. CEIs and study symbols are not assigned to hypothetical edges. Users can highlight edges that reflect the article’s main ideas, so that they are more apparent. In cases where no one relation has received dominant evidence, the corresponding edge is represented by a diamond arrowhead and is not assigned a CEI. © 2018 Matiasz et al. [MWD18]; licensed under CC BY 4.0.

3.2 Creating research maps: The local map

Figure 3.4 shows the interface for creating research maps. There are fields for: the what, where, and when properties for both the agent and the target; the study class; the type of result; and, for empirical edges, succinct descriptions of the approaches used to (1) observe or intervene on the agent and (2) measure changes in the target. When information is entered into the fields, the research map is updated accordingly. When a research map is created for an article that is indexed on PubMed, it is made public to all users. However, being first and foremost a tool for the personal

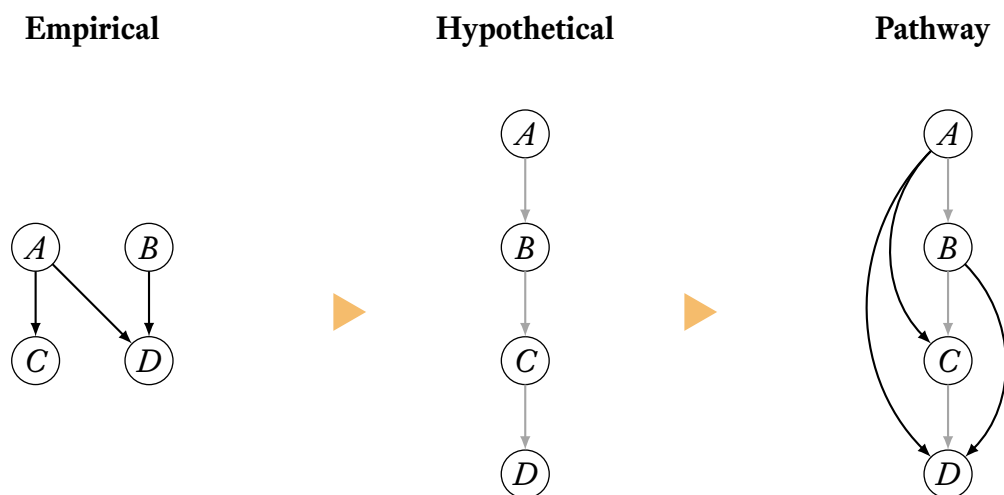


Figure 3.2: Using hypothetical edges to organize research maps. The diagram above shows how hypothetical edges (in gray) help to organize empirical edges in a research map, thus framing the empirical results in light of a specific hypothesis. Original figure © 2018 Matiasz et al. [MWD18]; used here under CC BY 4.0 with a different font.

curation of research information, ResearchMaps can also be used to create private maps, visible only to the user who entered them. These private maps can include unpublished experiments of ongoing projects, purely speculative models, etc.

There are multiple steps to make a research map for a given article. The first step is to identify all the nodes that will be included. This process entails the identification of agent–target pairs involved in the reported studies. For any one agent–target pair, the next step is to find the study class that was performed to test their relation. In addition to the study class, the user records the result that was obtained, as well as the key techniques that were used to observe (or manipulate) the agent and observe the result in the target. Once the empirical edges are entered, any hypothetical edges suggested by the article can be added, thereby helping to structure the map and contextualize the empirical results. Finally, because research maps can become large and complex, users can highlight the main edges, whether they are hypothetical or empirical.

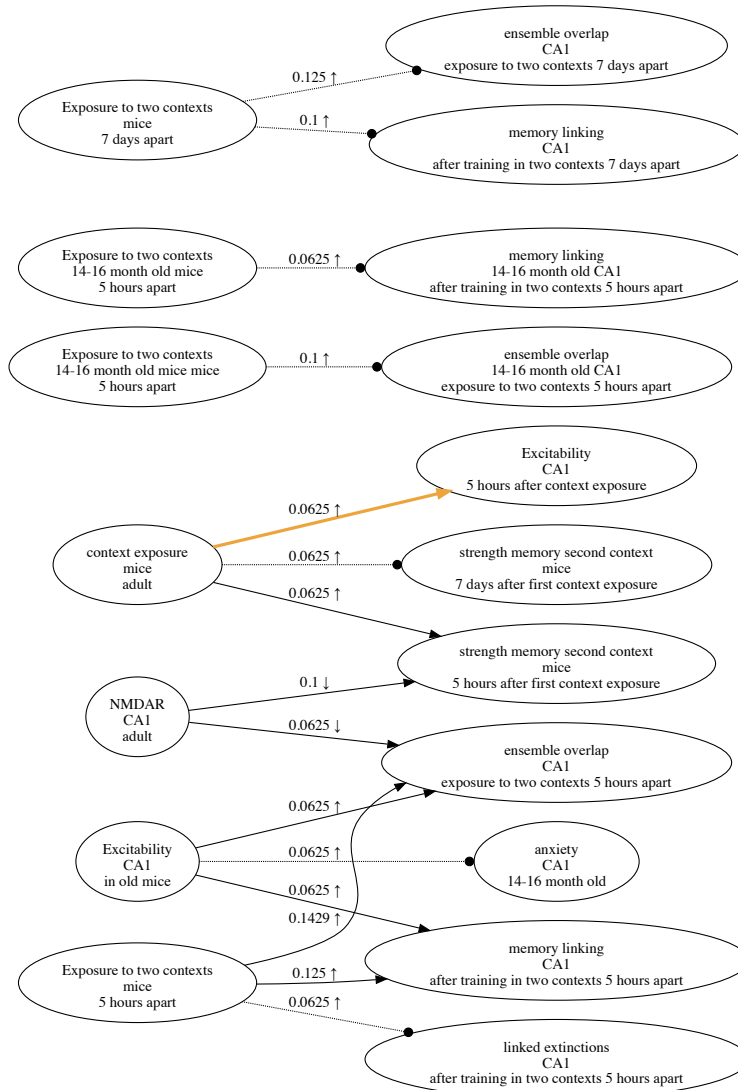


Figure 3.3: The research map of Figure 3.1 with its hypothetical edges removed. This modified research map, when compared with the one in Figure 3.1, illustrates how hypothetical edges help to structure research maps, thereby augmenting the interpretation of empirical results. © 2018 Matiasz et al. [MWD18]; licensed under CC BY 4.0.

3.3 Querying research maps: The global map

In addition to viewing the research maps of individual research articles, users can interact with all of the public data and their individual private data via the global map. On this page, users can search the application’s database either for a specific node (with a *What*, *Where*, and *When*) or simply for

a term—e.g., the transcription factor CREB (Figure 3.5). In addition to searching for a single entity, users can search for specific agent–target pairs, whether they are empirical or hypothetical. The queries operate on the union of all the (local) research maps that exist in the application’s database.

To constrain the visualizations produced by queries, users can modify each global search with several parameters, including a minimum and maximum threshold for filtering empirical edges based on their CEIs. By filtering out edges with low CEIs, for example, users can visualize only those edges with the highest levels of evidence—i.e., those that are likely to be more reliable. Similarly, by filtering out edges with high CEIs, users can quickly identify those connections with the least amount of evidence—i.e., those in greatest need of further investigation. Users can also limit the number of edges that must be traversed between a given query term and its results. Additionally, users can limit global searches to only the information that they personally entered, thus focusing searches to specific domains of interest. By interacting with the information in ResearchMaps, users can thus explore the ramifications of different hypotheses.

Clicking on any edge in the global map generates a table that lists all the empirical results and hypothetical assertions represented by that edge (Figure 3.6). Also provided are hyperlinks to the (local) research maps where this information was originally entered.

3.4 Data collected for analysis

Analyses presented in this dissertation use ResearchMaps data collected between 2013 and 2018. The bulk of this data was entered by the neuroscientist Alcino J. Silva; over three years, he created public research maps for 125 articles with 2,251 experiments, 1,293 nodes, and 1,693 edges. Figure 3.7 shows an aggregate of his research maps in memory allocation and other connected areas.

3.5 Details of the software implementation

The source code for ResearchMaps is publicly available at <https://github.com/ResearchMaps/>. The application is currently hosted by Amazon Web Services (AWS) Elastic Compute Cloud (EC2) with the Ubuntu 12.04 64-bit operating system.

ResearchMaps uses Node.js (<https://nodejs.org/>) as its runtime environment. HTML com-

ponents are made with the Bootstrap framework (<http://getbootstrap.com/>), and D3.js [BOH11] is used to modify the visualized research maps, which are created as SVG files with Graphviz [EGK01]. JavaScript and the jQuery library (<https://jquery.com/>) are also used. PubMed’s application programming interface (API) (<http://eutils.ncbi.nlm.nih.gov>) provides bibliographic information for published research articles, and the NeuroLex API, maintained by the Neuroscience Information Framework (NIF) [GAA08], provides suggested auto-completions for users’ input.

ResearchMaps uses the Neo4j 2.2.1 graph database and its query language Cypher. The database schema is designed as follows. Each user is assigned a `User` node, which is connected to `Paper` nodes that represent each research article (or private project) for which a user creates a research map. Each `Paper` node is connected to a number of `Experiment` nodes—one for each experiment (or hypothetical assertion) that is entered for a given map. Each `Experiment` node is connected to two `NeuroLexTerm` nodes representing the agent and the target for that particular experiment. Agent and target (`NeuroLexTerm`) nodes are connected by edges with properties to store the information used to calculate each edge’s CEI.

Mechanism for the learning deficits in a mouse model of neurofibromatosis type 1.

Costa RM, Federov NB, Kogan JH, Murphy GG, Stern J, Ohno M, Kucherlapati R, Jacks T, Silva AJ Departments of Neurobiology, Psychiatry and Psychology, BRI, University of California at Los Angeles, Los Angeles, California 90095-1761, USA.

Nature 2002 Jan 31;415(6871):526-30

[Search PubMed for this article](#)

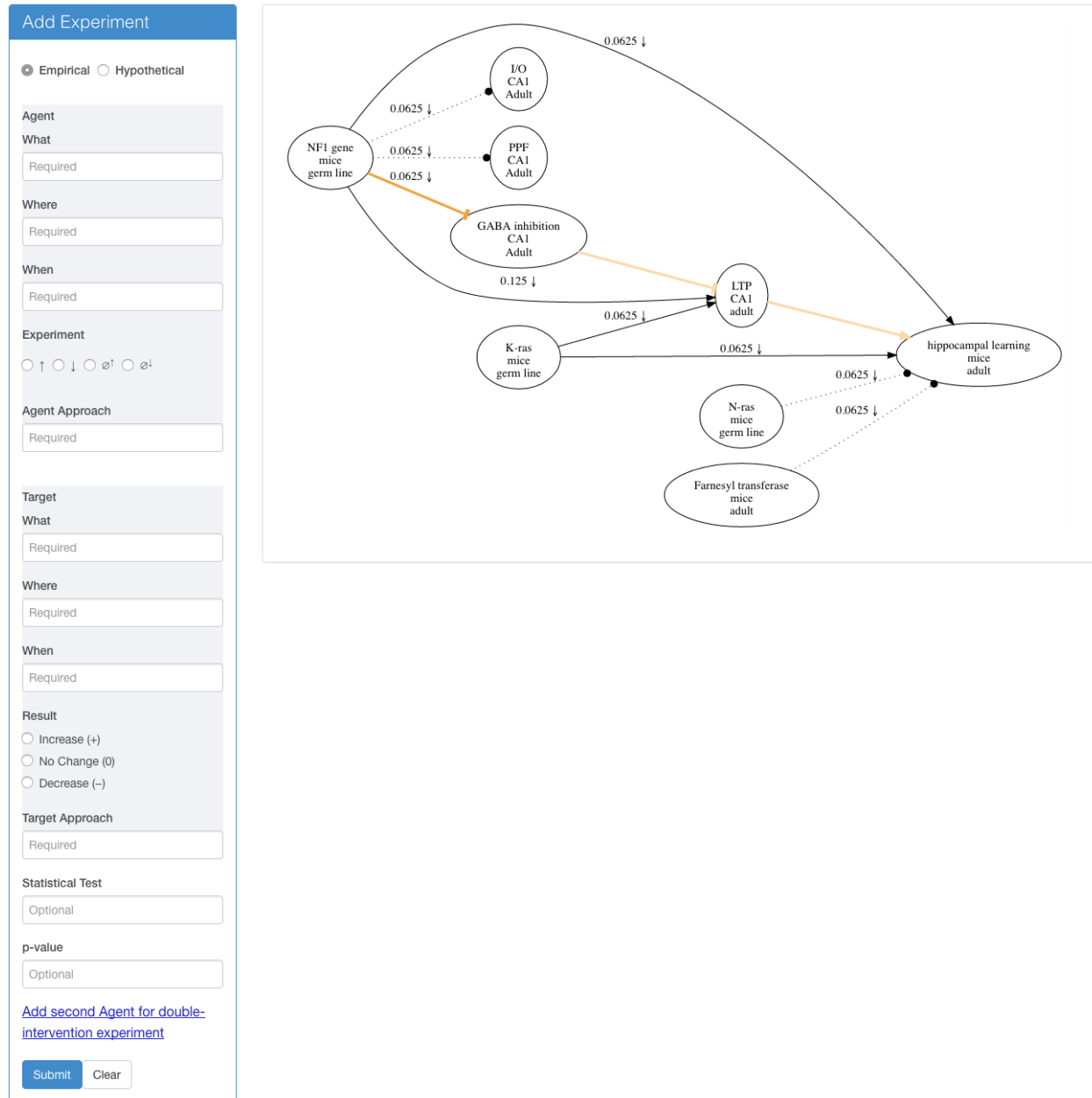


Figure 3.4: The local map of ResearchMaps. The form on the left is used to input information. The citation on the top indicates the article whose research map is displayed. Highlighted in yellow are edges that reflect the article’s main ideas. Users can double-click on any edge to retrieve PubMed citations that are potentially relevant to the agent–target relation. © 2018 Matiasz et al. [MWD18]; licensed under CC BY 4.0.

Global Map

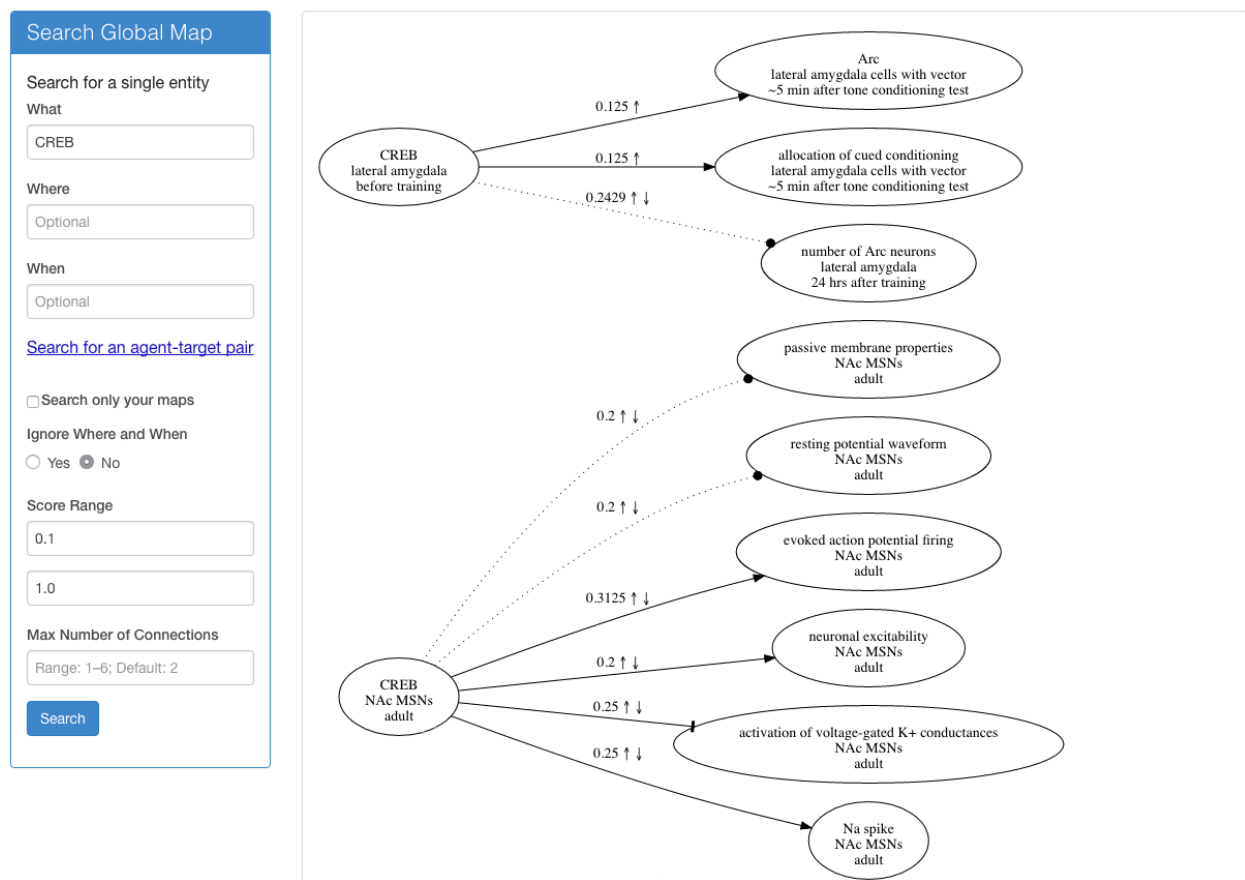


Figure 3.5: The global map of ResearchMaps. The form on the left is used to query all the research maps in the application’s database. On the right is a panel that displays the research map returned in response to the query. © 2018 Matiasz et al. [MWD18]; licensed under CC BY 4.0.

Global Map

Search Global Map

Search for a single entity

What

Where

When

[Search for an agent-target pair](#)

Search only your maps

Ignore Where and When

Yes No

Score Range

Max Number of Connections

Search

CREB

...

0.2429

↑ ↓

number of Arc neurons

Paper	What Agent	Where Agent	When Agent	Experiment	Agent Approach	What Target	Where Target	When Target	Result	Target Approach	Connection Type
View source map	CREB	lateral amygdala	before training	Negative	HSV viral vector with mutant CREBS133A gene in about 15% of the cells	number of Arc neurons	lateral amygdala	24 hrs after training	No Change	CATFISH in animals after training with 0.4 mA US	No Relation
View source map	CREB	lateral amygdala	before training	Positive	HSV viral vector with wild-type CREB gene in about 15% of the cells	number of Arc neurons	lateral amygdala	24 hrs after training	No Change	CATFISH in animals after training with 0.4 mA US	No Relation
View source map	CREB	lateral amygdala	before training	Positive	HSV viral vector with wild-type CREB gene in about 15% of the cells	number of Arc neurons	lateral amygdala	24 hrs after training	No Change	CATFISH in animals after training with 0.75 mA US	No Relation
View source map	CREB	lateral amygdala	before training	Negative	HSV viral vector with mutant CREBS133A gene in about 15% of the cells	number of Arc neurons	lateral amygdala	24 hrs after training	No Change	CATFISH in animals after training with 0.4 mA US	No Relation
View source map	CREB	lateral amygdala	before training	Positive	HSV viral vector with wild-type CREB gene in about 15% of the cells	number of Arc neurons	lateral amygdala	24 hrs after training	No Change	CATFISH in animals after training with 0.75 mA US	No Relation
View source map	CREB	lateral amygdala	before training	Positive	HSV viral vector with wild-type CREB gene in about 15% of the cells	number of Arc neurons	lateral amygdala	24 hrs after training	No Change	CATFISH in animals after training with 0.4 mA US	No Relation

Figure 3.6: Provenance of edges in the global map. Each edge in the global map can be clicked, revealing a table that lists every empirical result or hypothetical assertion recorded for that edge. Each entry in this table has a link to the local research map that contains the edge that was clicked.

© 2018 Matiasz et al. [MWD18]; licensed under CC BY 4.0.

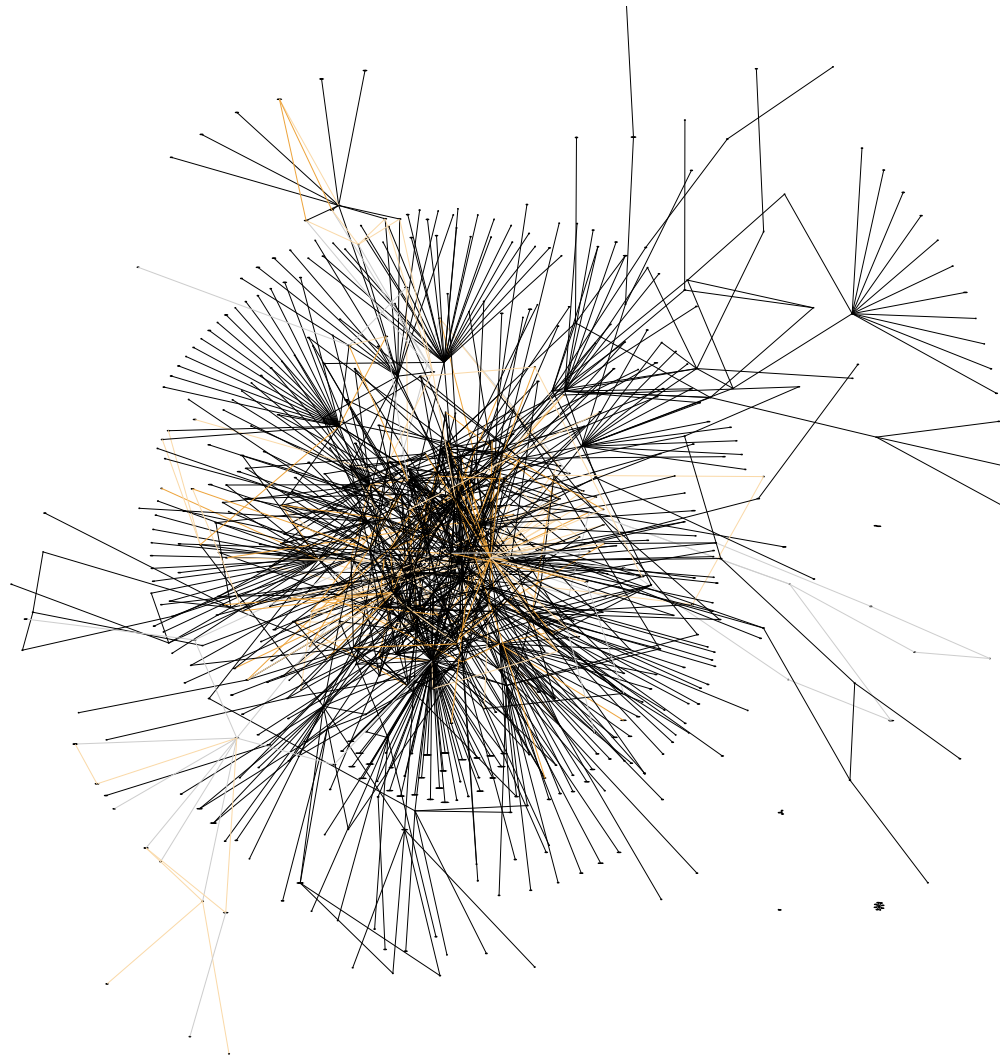


Figure 3.7: This is Alcino Silva’s personally curated research map of work in the field of memory allocation, as well as related work that either overlaps or connects to the work in memory allocation. To minimize the number of nodes, only the *What* property of each node is shown, so that nodes with different *Where* and *When* properties (but identical *What* properties) are collapsed into one. Nodes in orange appear only in research maps for articles on memory allocation. Nodes in red appear not only in research maps for articles on memory allocation but also in research maps of related work.
© 2018 Matiasz et al. [MWD18]; licensed under CC BY 4.0.

CHAPTER 4

Collecting constraints on causal structure

Data that is collected from a system can be analyzed to identify conditional dependence and independence relations among the system’s variables. These statistical relations can then be used to infer the causal structure that governs the system from which the data was collected. This inference is permitted by *bridge principles*, which “connect what can be observed to the underlying causal structure that generates the phenomena” [Ebe09]. In this dissertation, what can be observed is expressed as conditional (in)dependence relations; the underlying causal structure is expressed as a causal graph. The bridge principles that connect (in)dependence relations to causal graphs are the causal Markov and causal faithfulness conditions (§ 2.2). Together, these conditions allow for a relation between conditional (in)dependencies in a probability distribution and specific structures in a causal graph [SGS00]. For instance, if two variables in a system are statistically dependent, the system’s causal graph will have certain features, such as one or more paths that correspond to this statistical dependence. The (in)dependence relations obtained from a system thus constrain which causal structures can accurately describe the system, and such relations can serve as inputs to constraint-based causal discovery algorithms (e.g., [HHE13, HEJ14]).

We express causal-structure constraints in the form $X \perp\!\!\!\perp Y \mid \mathbf{C} \parallel \mathbf{J}$, where X and Y are two variables involved in an independence relation; \mathbf{C} is a (possibly empty) set of variables on which we must statistically condition for the relation to hold; and \mathbf{J} is a (possibly empty) set of variables that underwent experimental intervention when the relation manifested [HEJ14]. Dependence statements instead use the “*not*-independent” symbol ($\not\perp$). The empty-set symbol (\emptyset) is used to denote empty sets for \mathbf{C} and \mathbf{J} . An example of an (in)dependence relation is

long-term potentiation $\not\perp$ spatial learning $\mid \emptyset \parallel$ long-term potentiation,

which states that long-term potentiation and spatial learning were observed to be (unconditionally)

dependent in an experiment that intervened on long-term potentiation.

Below, § 4.1 describes how research-map annotations of the literature can be translated into formal constraints on causal structure for use in constraint-based causal discovery, which is presented in Chapter 5. I refer the reader to Eberhardt [Ebe17] for a discussion of how one can formalize background assumptions to further constrain the search over causal structures.

4.1 Annotating empirical results in literature

Although the primary data for many studies remains inaccessible to most researchers, research articles commonly report statistical information that can be formalized as constraints on causal structure. Examples of this information include statistical tests for the correlation between measured variables. A research map that captures this statistical information can thus be translated into (in)dependence relations, to be used for constraint-based causal discovery. Table 4.1 presents this translation for research map annotations involving two phenomena, one agent and one target. In this translation, a research map’s distinction between a positive correlation (an excitatory relation) and negative correlation (an inhibitory relation) is discarded in the causal-structure constraint: these cases are both mapped to a dependence relation (\perp). (In § 9.5.5, I discuss the possibility of incorporating this sign information into the search over causal structures.)

A research map is not the only representation that can be used to represent causal-structure constraints from the literature. In principle, one could use any representation with all the components required to instantiate an (in)dependence relation. Research maps were used for the work presented in this dissertation primarily because the ResearchMaps web application greatly facilitates this annotation.

These annotations can be used as input for constraint-based causal discovery, yielding causal graphs that are maximally consistent with the evidence that has been extracted from literature. This process, which is described in the next chapter, formalizes what scientists normally do when they read research articles: they attempt to “stitch” the various findings together, forming a coherent picture of a system whose dynamics are consistent with the annotated evidence. But as I discuss in § 1.1, a body of evidence rarely identifies one unique causal graph; instead, it is usually consistent with multiple graphs, each with its own causal structure. Through their ability to identify *every*

Method	Agent's change	Target's change	Relation	Constraint
intervention	increase	increase	excitatory	$A \not\perp T \mid \emptyset \parallel A$
		no change	no-connection	$A \perp T \mid \emptyset \parallel A$
		decrease	inhibitory	$A \not\perp T \mid \emptyset \parallel A$
	decrease	increase	inhibitory	$A \not\perp T \mid \emptyset \parallel A$
		no change	no-connection	$A \perp T \mid \emptyset \parallel A$
		decrease	excitatory	$A \not\perp T \mid \emptyset \parallel A$
observation	increase	increase	excitatory	$A \not\perp T \mid \emptyset \parallel \emptyset$
		no change	no-connection	$A \perp T \mid \emptyset \parallel \emptyset$
		decrease	inhibitory	$A \not\perp T \mid \emptyset \parallel \emptyset$
	decrease	increase	inhibitory	$A \not\perp T \mid \emptyset \parallel \emptyset$
		no change	no-connection	$A \perp T \mid \emptyset \parallel \emptyset$
		decrease	excitatory	$A \not\perp T \mid \emptyset \parallel \emptyset$

Table 4.1: The translation of research map annotations to (in)dependence relations for use in constraint-based causal discovery. This table includes research map annotations involving only one agent and one target.

consistent causal graph, causal discovery algorithms help scientists to avoid bias when they search for consistent graphs and construct new hypotheses based on them. Evaluations in § 8.3 show how research-map annotations can prune the viable model space of causal graphs, bringing scientists closer to the true causal graph.

CHAPTER 5

Identifying consistent causal structures

This chapter presents the constraint-based causal discovery algorithm introduced by Hyttinen et al. [HEJ14]. Although this method is not a contribution of this dissertation, it is a crucial component of the meta-analytic pipeline presented in Figure 1.4 and is thus presented here for completeness. Technical details regarding the software implementation of the algorithm are also provided. The source code for this method is currently available at <https://sites.google.com/site/ajhyttin/>. This algorithm was chosen for the pipeline because it is currently the state of the art in causal discovery. Among current methods, it considers the most general model space: neither acyclicity nor causal sufficiency needs to be assumed; the algorithm can thus consider models that contain both cycles (feedback) and latent confounders. Additionally, the algorithm’s constraint-based approach enables the formalization of background assumptions [Ebe17], as well as the degree-of-freedom approach described in § 6.2.

The intuition for this algorithm is as follows. Scientists will perform experiments to understand the causal relations that govern the phenomena in a system. These phenomena and the causal relations between them can be represented by the nodes and directed edges that compose a causal graph. We will call this causal graph that correctly models the system the *true* causal graph. In addition to this true graph, there are other graphs with the same variables but different sets of edges, corresponding to different causal explanations of the system’s behavior. The number of possible causal graphs is very large, even for small sets of variables (§ 2.2). Thus, the scientist who performs experiments to identify the true causal graph is “searching for a needle in a really huge haystack of falsehoods” [Gly04].

An experiment’s result can show the scientists which parts of the haystack are safe to remove: namely, all the causal graphs that are inconsistent with the result.¹ When a result is expressed as

¹ An erroneous result can mislead scientists by motivating them to remove a part of the haystack that in fact contains

a conditional (in)dependence relation (Chapter 4), the rules of d-separation (§ 2.2) can be used to identify the particular causal graphs that are consistent with the result. Any scientist who understands d-separation can use a pen and paper to check whether an (in)dependence relation is consistent with a causal graph. But this computation is infeasible to do manually when there are thousands of possible graphs, as is true even for a system with only five variables. Therefore, the strategy taken by Hyttinen et al. is to have this computation performed by a machine.

The algorithm uses answer set programming (ASP), a type of logic programming that is useful for solving very challenging problems such as NP-hard optimization tasks. It is based on the concept of declarative constraint satisfaction [GL88, Bar03]. In this context, the constraints are (in)dependence relations, and they are satisfied only by the particular causal graphs that encode those relations, as given by the rules of d-separation.

The algorithm proceeds in the following steps. First, (in)dependence relations among the system’s variables are obtained—either by performing statistical independence tests on data [HEJ14], or by annotating qualitative information in the literature (§ 2.1 and § 4.1). If none of the constraints conflict with each other, then a Boolean satisfiability (SAT) solver [BHM09] is sufficient to find the consistent causal graphs [HHE13]. However, if the constraints contain conflicts—for instance, if one constraint states that X and Y are independent, while another states that they are dependent—then a Boolean *maximum* satisfiability (MaxSAT) solver [BHM09] is required: in this case, each constraint is assigned a weight that denotes its confidence, and the solver finds the causal graphs that minimize the sum of the weights for unsatisfied constraints. Weights can be assigned based on the p -values of independence tests [HEJ14] or based on other measures of confidence, such as the cumulative evidence index (§ 6.1) for the research-map edge from which the constraint was derived (§ 4.1). Hyttinen et al. present the constrained optimization problem as follows: given a set \mathbf{K} of conditional (in)dependence constraints for a set of variables \mathbf{V} , and a non-negative weight $w(k)$ for each k in \mathbf{K} , we wish to find the causal graph G^* (from the class of causal graphs, \mathcal{G} , with vertices in \mathbf{V}) such that

$$G^* \in \operatorname{argmin}_{G \in \mathcal{G}} \sum_{k \in \mathbf{K}: G \not\models k} w(k). \quad (5.1)$$

the needle (i.e., the true causal graph). This dissertation does not model scientists’ fallibility; instead, the focus is on how to reason with evidence and plan experiments, assuming that those experiments will be performed competently.

A state-of-the-art MaxSAT solver named Clingo [GKK11] is guaranteed to converge to a globally optimal solution, thus identifying the causal graphs that maximally satisfy the constraints.

Clingo reads independence constraints in the form $\text{indep}(X, Y, C, J, M, W)$; dependence constraints take the form $\text{dep}(X, Y, C, J, M, W)$. For both forms, X and Y are two variables involved in an (in)dependence relation; C denotes the conditioning set C ; and J denotes the intervention set J (Chapter 4). The parameter M denotes the marginalization set; for all the analyses and simulations in this dissertation, the set M is equal to $V \setminus (\{X, Y\} \cup C \cup J)$, which is the set of all variables in the system other than the variables in $\{X, Y\}$, C , and J . Lastly, the parameter W denotes the weight assigned to the constraint.

Variables X and Y are indexed according to an integer index $(1, 2, 3, \dots)$. For example, a system with four variables would assign the indices 1, 2, 3, and 4 to the variables. Every constraint lists the parameters X and Y in ascending order. For instance, if variables 1 and 3 appear in a dependence relation, the constraint is written $\text{dep}(1, 3, \dots)$, *not* as $\text{dep}(3, 1, \dots)$. Unlike X and Y , the parameters C , J , and M are indexed according to the following binary scheme: given a system with N variables, a string of N binary digits is used as a set of indicator variables to show which of the system's variables are included in a given set. For example, if $N = 4$, and we want to construct a conditioning set that contains only the variable with the integer index 4, we construct the string of digits 1000 in binary notation, which is equivalent to 8 in decimal notation. In this case, parameter C in the constraint syntax would be set to 8, yielding the constraint $\text{dep}(X, Y, 8, \dots)$. As another example, if we want to construct an intervention set with the integer indices 3 and 4, we construct the string 1100 in binary and thus 12 in decimal. Parameter J would thus be set to 12, yielding the constraint $\text{dep}(X, Y, C, 12, \dots)$. Table 5.1 gives the ASP-encoded constraints for the (in)dependence relations in Figure 1.4.

A set of (in)dependence constraints is input into Clingo, which then computes the causal graphs that are optimal according to Equation 5.1. Clingo can output either *one* optimal solution or *every* optimal solution; the latter case is invoked by adding the flag `-opt-mode=optN` in the call to Clingo [GKK15]. When there are no conflicting constraints, all the output graphs will be consistent with all the input constraints, yielding a Markov equivalence class (§ 2.2.1). When there are conflicting constraints, the graphs will not be consistent with all the constraints, but they will

(In)dependence relation	ASP input
$X \not\perp\!\!\!\perp Y \mid \emptyset \parallel \emptyset$	<code>dep(1,2,0,0,4,W)</code>
$Y \not\perp\!\!\!\perp Z \mid \emptyset \parallel \emptyset$	<code>dep(2,3,0,0,1,W)</code>
$X \not\perp\!\!\!\perp Z \mid \emptyset \parallel \emptyset$	<code>dep(1,3,0,0,2,W)</code>
$X \perp\!\!\!\perp Z \mid \{Y\} \parallel \emptyset$	<code>indep(1,3,2,0,0,W)</code>

Table 5.1: The ASP encodings for the (in)dependence relations (i.e., Clingo’s input) in Figure 1.4. The variables X , Y , and Z are identified with the integer indices 1, 2, and 3, respectively. Weights are not assigned to these constraints; the parameter W is used as a placeholder.

still be optimal according to Equation 5.1. As a shorthand, this dissertation also refers to this set of graphs as an equivalence class (§ 2.2.1 explains these two usages of “equivalence class”).

Each causal graph in Clingo’s output is described by set of statements of the form `edge(X, Y)`. For example, the integer-indexed graph $1 \rightarrow 2 \rightarrow 3$ is encoded by the statements `edge(1, 2)` and `edge(2, 3)`. The graph $1 \leftarrow 2 \rightarrow 3$ is encoded by `edge(2, 1)` and `edge(2, 3)`. If Clingo is run without assuming causal sufficiency, graphs with latent variables (and thus confounding relations) are encoded with the syntax `conf(X, Y)`, which denotes that the variables X and Y are confounded. Table 5.2 uses this encoding to express the equivalence class in Figure 1.4.

Causal graph	ASP output
$X \rightarrow Y \rightarrow Z$	<code>edge(1,2), edge(2,3)</code>
$X \leftarrow Y \rightarrow Z$	<code>edge(2,1), edge(2,3)</code>
$X \leftarrow Y \leftarrow Z$	<code>edge(2,1), edge(3,2)</code>

Table 5.2: The ASP encoding (i.e., Clingo’s output) for the causal graphs in Figure 1.4. The variables X , Y , and Z are identified with the integer indices 1, 2, and 3, respectively.

Scientists can use algorithms like the one above to find causal explanations for their data. Thanks to the mathematization of causality, this task can now be performed objectively and exhaustively, diminishing the potential for bias. Scientists can spend their time on more challenging tasks that have not yet been operationalized with such precision, such as designing experimental

protocols and defining the phenomena that are to be included in a causal model. This reliance on machines to find consistent causal graphs should be no more taboo than the reliance on machines to compute descriptive statistics for datasets. What is true of arithmetic is now true of this causal-reasoning task: what once needed to be done by hand should now be performed more efficiently by machines. Of course, there can always be disagreements regarding modeling assumptions, but these considerations also apply to causal reasoning performed by humans. In the next chapter, I discuss how machine-readable data structures like causal graphs can be used as the basis for experiment planning, which entails causal reasoning. Although designing an experiment has not been formalized to the same extent, aspects of this process can be made more objective and communicable by using causal graphs as the representational tool over which we quantify causal underdetermination. The use cases presented in § 8.2 illustrate how human cognition can be augmented by these computational approaches.

CHAPTER 6

Quantifying evidence and causal underdetermination

When scientists amalgamate empirical results, they can view the activity as trying to quantify either evidence or uncertainty. This chapter shows how these two complementary approaches can be guided by research maps and causal graphs.

6.1 Quantifying evidence in research maps

Evidence is meaningful only when it entails a relation between facts and hypotheses. Facts alone cannot constitute evidence whose weight or strength can be measured, devoid of context [Goo50, OG74, VC18]. For instance, a patient’s cough is not stronger evidence than his bruise until these facts are considered with respect to the hypothesis that the patient is sick. If the hypothesis changes, the weight of the evidence changes accordingly: if we hypothesize that the patient has fallen, his bruise now carries more evidential weight than his cough.

The research-map framework defines a *cumulative evidence index* (CEI) for quantifying evidence [MWD18]. In a research map, each fact is an empirical result regarding two phenomena; each hypothesis posits a specific relation between these two phenomena—either excitation, inhibition, or no-connection (§ 2.1). The CEI quantifies the extent to which empirical results lend evidence to these relations. This calculation uses a Bayesian model of scientific consensus building to express many of the commonsense intuitions that researchers use to reason about evidence.

The CEI for a research-map edge is calculated as follows. Let $C = \{\uparrow, \emptyset^\uparrow, \emptyset^\downarrow, \downarrow\}$ denote the set of all study classes, where $c = \uparrow$ denotes the class positive intervention; $c = \emptyset^\uparrow$ denotes the class positive non-intervention; $c = \emptyset^\downarrow$ denotes the class negative non-intervention; and $c = \downarrow$ denotes the class negative intervention. Let $R = \{\mathcal{E}, \mathcal{N}, \mathcal{I}\}$ denote the set of relations that can exist between two phenomena and for which a study can provide evidence, where \mathcal{E} denotes an excitatory

relation; \mathcal{N} denotes a no-connection relation; and \mathcal{I} denotes an inhibitory relation. Thus, a study of class $c \in \{\uparrow, \emptyset^\uparrow, \emptyset^\downarrow, \downarrow\}$ can yield evidence in support of relation $r \in \{\mathcal{E}, \mathcal{N}, \mathcal{I}\}$.

Let $\alpha_c = (\alpha_{c,\mathcal{E}}, \alpha_{c,\mathcal{N}}, \alpha_{c,\mathcal{I}})$; let $\theta_c = (\theta_{c,\mathcal{E}}, \theta_{c,\mathcal{N}}, \theta_{c,\mathcal{I}})$, and let $x_c = (x_{c,\mathcal{E}}, x_{c,\mathcal{N}}, x_{c,\mathcal{I}})$, where

$$(\theta_{c,\mathcal{E}}, \theta_{c,\mathcal{N}}, \theta_{c,\mathcal{I}}) \sim \text{Dir}(\alpha_{c,\mathcal{E}}, \alpha_{c,\mathcal{N}}, \alpha_{c,\mathcal{I}}), \quad (6.1)$$

$$(x_{c,\mathcal{E}}, x_{c,\mathcal{N}}, x_{c,\mathcal{I}}) \sim \text{Mult}(\theta_{c,\mathcal{E}}, \theta_{c,\mathcal{N}}, \theta_{c,\mathcal{I}}, n_c). \quad (6.2)$$

Here, $\alpha_{c,r}$ is the prior weight given to relation r supported by studies of class c ; $\theta_{c,r}$ is the probability that the next study of class c will yield evidence in support of relation r ; $x_{c,r}$ is the number of studies of class c that have yielded evidence in support of relation r ; and n_c is the number of studies of class c that have been performed. For each study class c , we can define x_c (compare to the table in Figure 6.1):

$$x_\uparrow = [x_{\uparrow,\mathcal{E}}, x_{\uparrow,\mathcal{N}}, x_{\uparrow,\mathcal{I}}], \quad (6.3)$$

$$x_{\emptyset^\uparrow} = [x_{\emptyset^\uparrow,\mathcal{E}}, x_{\emptyset^\uparrow,\mathcal{N}}, x_{\emptyset^\uparrow,\mathcal{I}}], \quad (6.4)$$

$$x_{\emptyset^\downarrow} = [x_{\emptyset^\downarrow,\mathcal{E}}, x_{\emptyset^\downarrow,\mathcal{N}}, x_{\emptyset^\downarrow,\mathcal{I}}], \quad (6.5)$$

$$x_\downarrow = [x_{\downarrow,\mathcal{E}}, x_{\downarrow,\mathcal{N}}, x_{\downarrow,\mathcal{I}}]. \quad (6.6)$$

The CEI for an edge is based on the values of θ_c for each of the study classes, which are updated as additional studies are recorded, thereby changing the values of x_c . We are thus interested in estimating each θ_c in light of the evidence represented by each x_c . Applying Bayes' theorem yields:

$$p(\theta_c | x_c, \alpha_c) \propto p(x_c | \theta_c) p(\theta_c | \alpha_c), \quad (6.7)$$

$$\propto \theta_{c,\mathcal{E}}^{\alpha_{c,\mathcal{E}} + x_{c,\mathcal{E}} - 1} \theta_{c,\mathcal{N}}^{\alpha_{c,\mathcal{N}} + x_{c,\mathcal{N}} - 1} \theta_{c,\mathcal{I}}^{\alpha_{c,\mathcal{I}} + x_{c,\mathcal{I}} - 1}, \quad (6.8)$$

The posterior distribution is in the form of a Dirichlet distribution, so we have that:

$$\theta_c | x_c, \alpha_c \sim \text{Dir}(\alpha_c + x_c). \quad (6.9)$$

The expected value of this distribution is thus expressed as:

$$E[\theta_{c,r} | x_c, \alpha_c] = \frac{\alpha_{c,r} + x_{c,r}}{\sum_r \alpha_{c,r} + n_c}. \quad (6.10)$$

If $\alpha_{c,r} = 1$ for all c and r , the above expression becomes:

$$E[\theta_{c,r} | x_{c,r}, \alpha_{c,r} = 1] = \frac{1 + x_{c,r}}{|R| + n_c}, \quad (6.11)$$

which is an implementation of Laplace (add-one) smoothing.¹

In the absence of evidence (i.e., before any studies are performed), $x_{c,r} = 0$ for all c, r . We denote this state by θ_o :

$$\theta_o = E[\theta_{c,r} | x_c = (0,0,0), \alpha_{c,r} = 1] = \frac{1}{|R|} = \frac{1}{3}. \quad (6.12)$$

Let $\bar{\theta}$ denote the set of mean r -components across all the study classes—an expression of convergence:

$$\bar{\theta} = \frac{1}{|C|} \left[\sum_c E[\theta_{c,\mathcal{E}} | x_{c,\mathcal{E}}, \alpha_{c,\mathcal{E}} = 1], \sum_c E[\theta_{c,\mathcal{N}} | x_{c,\mathcal{N}}, \alpha_{c,\mathcal{N}} = 1], \sum_c E[\theta_{c,\mathcal{I}} | x_{c,\mathcal{I}}, \alpha_{c,\mathcal{I}} = 1] \right]. \quad (6.13)$$

The relation assigned to the research-map edge is the relation with the largest component in $\bar{\theta}$:

$$\operatorname{argmax}_r \bar{\theta}_r. \quad (6.14)$$

The CEI assigned to the research-map edge equals:

$$\frac{\max \bar{\theta} - \theta_o}{1 - \theta_o}, \quad (6.15)$$

where $\max \bar{\theta}$ denotes the largest component of $\bar{\theta}$. In cases where two or more components of $\bar{\theta}$ are equal, neither a relation nor a CEI is assigned to the edge.

To develop an intuition for this scoring approach, consider the following example, which uses the studies involving CREB and the number of Arc neurons that are depicted in Figure 6.2. In this research map, the edge connecting these two nodes represents three studies: two positive interventions of CREB resulting in no change in the number of Arc neurons, and one negative intervention of CREB, again resulting in no change. Together, these three studies provide evidence for a no-connection edge between the two nodes. Before any of these studies were performed, θ_c was uniform for all c . After the first study, in which a positive intervention produced no change in the target, $\theta_c = (0.25, 0.50, 0.25)$ and the CEI of the edge was 0.0625. After the second positive intervention (with the same result as the first), the CEI of the edge became 0.1000.

¹ I am grateful to Justin Wood for suggesting that we use Laplace smoothing.

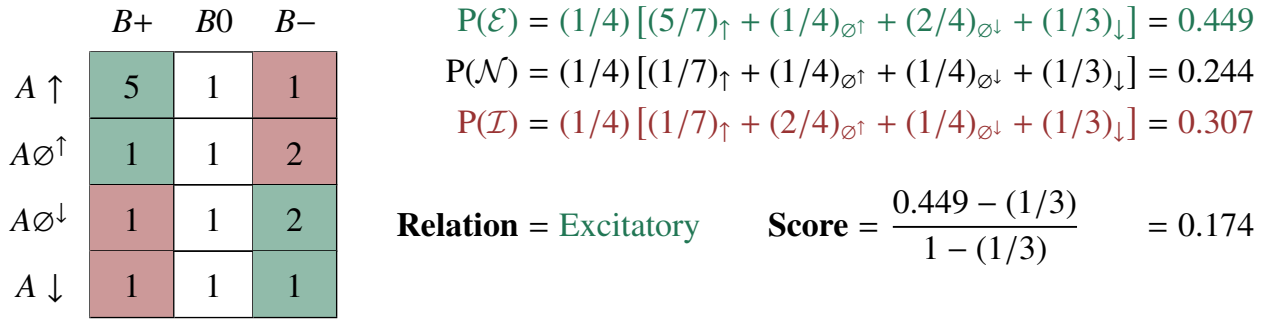


Figure 6.1: A shorthand method for calculating the CEI for an edge in a research map. A table representing the model space of studies is instantiated with a pseudocount of one (a form of Laplace smoothing). The symbols along the left indicate the study classes involving an agent, A : positive intervention ($A \uparrow$), positive non-intervention ($A \emptyset \uparrow$), negative non-intervention ($A \emptyset \downarrow$), and negative intervention ($A \downarrow$). The symbols along the top indicate the results recorded in a target, B : increase ($B+$), no change ($B0$), and decrease ($B-$). This particular instantiation of the scoring table encodes four ($5 - 1$) positive interventions in which the target increased, one ($2 - 1$) positive non-intervention in which the target decreased, and one ($2 - 1$) negative non-intervention in which the target decreased. There are thus five studies suggesting an excitatory relation (green regions), and one study suggesting an inhibitory relation (red region). © 2018 Matiasz et al. [MWD18]; licensed under CC BY 4.0.

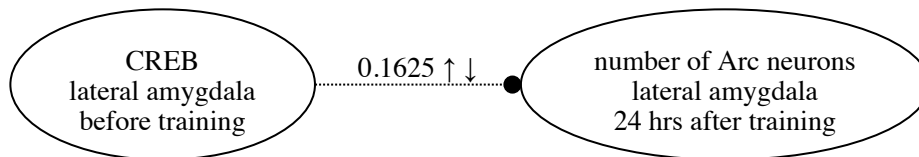


Figure 6.2: An example of an edge in a research map. This research map encodes three studies: two positive interventions (\uparrow) and one negative intervention (\downarrow). This edge is part of the research map for Han et al. [HKY07]. © 2018 Matiasz et al. [MWD18]; licensed under CC BY 4.0.

The first positive intervention thus changed the CEI by 0.0625, while the second experiment changed the CEI by 0.0375. These two changes in the CEI illustrate how the principle of consistency (§ 2.1) is expressed quantitatively by the scoring algorithm: each subsequent study that yields consistent results will increase the CEI, albeit by an amount that is less than the amount contributed by the previous consistent study in the same class.

After the third experiment, in which a previously unrepresented study class (negative intervention) yielded a consistent result (no change), the CEI increased to 0.1625, for a net change of 0.0625. This change demonstrates another desirable feature of the CEI: when consistent results are obtained across multiple study classes, each sequence of studies within a class contributes the same set of decaying amounts to the CEI, such that results across the four study classes are weighted independently of the order in which they are obtained.

If a fourth study with conflicting evidence were recorded—e.g., a positive non-intervention yielding an increase in the target—the CEI would drop to 0.1313. Appropriately, the conflicting evidence would undermine the still dominant evidence that the relation between the two nodes is no-connection. Had this conflicting evidence come from another positive intervention, a study class already represented in the CEI, the CEI would have dropped to 0.1250. This larger drop (compared to the one incurred for a conflicting positive non-intervention) reflects the principle of *divergence*, which is a corollary to the principle of convergence (§ 2.1): scientists tend to trust evidence from a particular study class to the extent that studies within this class yield consistent results; equivalently, scientists will mistrust a particular study class to the extent that studies within this class yield conflicting results.

This model can be tuned along multiple parameters; the particular values in the example above are specific to modeling assumptions that can be modified as needed. For instance, the model can accommodate any number of distinct study classes and relation types. Additionally, each study does not have to contribute an equal weight to the CEI: aspects of each study, such as its p -value or sample size, can be used to scale that study’s contribution. Similarly, each study class does not have to be weighted equally: if a particular field is known to value one study class over another, the weighting can reflect this preference, allowing the model to quantitatively express any *evidence hierarchy* [MDR18]. Finally, the rate at which the CEI approaches one—or, equivalently, the impact of the prior distribution—can be adjusted to change the number of studies that are needed to reach a high level of confidence.

In the next chapter, I discuss how a research map can help to plan experiments. A research map frames this task as the search for a study that could maximize the CEI for a single edge, or the CEIs for a set of edges. The CEI has not yet been generalized to allow for a single map-level CEI.

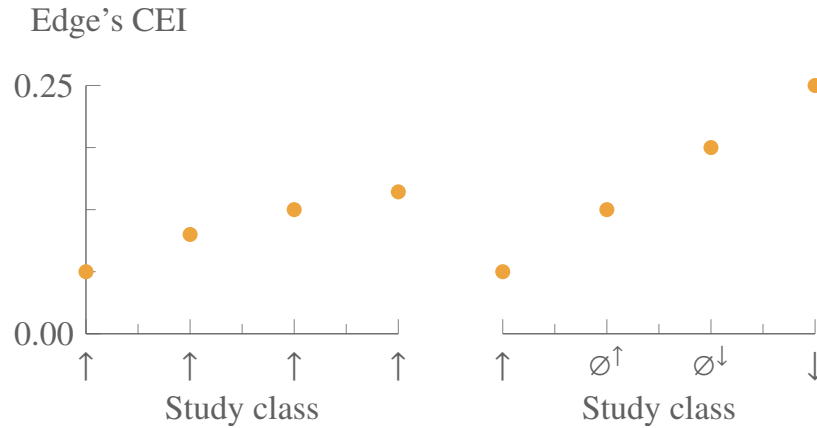


Figure 6.3: The growth of an edge’s CEI due only to consistency (left) and due to convergence (right). These plots show how the CEI of a research-map edge increases with each subsequent experiment (all with agreeing results), due to the principle of consistency (left) and due to the principle of convergence (right). The plot on the left represents repeated iterations of the same class of experiment (e.g., positive intervention) with consistent results. The plot on the right represents multiple iterations of experiments in which, at each iteration, one of the least-represented classes of experiments was performed, leading to consistent results. These two plots express an axiom of research maps: convergence carries greater epistemological weight than consistency. Original figure © 2018 Matiasz et al. [MWD18]; used here under CC BY 4.0 with modified axis labels.

But its semantics can be applied to multiple edges, yielding qualitative principles for experiment planning that could guide the formulation of a CEI that incorporates evidence from multiple edges.

6.2 Quantifying causal underdetermination in causal graphs

An equivalence class of causal graphs represents the range of causal interpretations that one can defensibly take in light of the available evidence. The diversity of causal structures in an equivalence class represents the extent to which the available evidence is lacking and the extent to which the true causal graph is *underdetermined*: the less evidence there is, the more causal graphs there are that remain consistent with what is known. Because this lack of knowledge is what drives scientific inquiry, quantifying a causal graph’s underdetermination can help scientists to determine which next experiments could be most instructive. We can quantify this underdetermination by considering the

diversity of causal structures that exist throughout all the graphs in an equivalence class.

The *degrees of freedom* for a causal graph are the possible variations in edge relations that can exist between any two variables throughout an equivalence class [MWW17b]. For DAGs, these edge relations are:

- a “left-to-right” edge ($X \rightarrow Y$);
- a “right-to-left” edge ($X \leftarrow Y$); and
- neither edge ($X \quad Y$).²

When we allow for cycles, there is a fourth relation consisting of both directed edges ($X \rightleftharpoons Y$). Here, we consider only the three edge relations for DAGs. To fully specify a causal graph over N variables, we need to instantiate exactly one of these edge relations for each of the $\binom{N}{2}$ pairs of variables in the graph. Once a particular edge relation is instantiated for a pair of variables (e.g., $X \rightarrow Y$), there are two other possible edge relations—two degrees of freedom—that the pair can take (e.g., $X \leftarrow Y$ and $X \quad Y$). The trivial equivalence class that contains every possible causal graph (satisfying zero constraints) thus has $2^{\binom{N}{2}}$ degrees of freedom. Note that this number is much smaller than the number of possible causal graphs over the same number of variables.

Each causal graph in an equivalence class instantiates these edge relations differently for at least one of the pairs of variables. For each pair of variables in a system, we can determine the number of instantiations that remain underdetermined by looking at the set of all edge relations that appear in the system’s equivalence class. For example, in the equivalence class of Figure 2.3, the graphs all agree that there is no edge for the pair $\{X, Z\}$. This edge relation is thus fixed: regardless of which graph is correct, we know what the edge relation for this pair is $X \quad Z$. The graphs in this equivalence class unanimously agree regarding the *existence* of edges for the pairs $\{X, Y\}$ and $\{Y, Z\}$; however, they do not unanimously agree regarding the edges’ *orientations*. This equivalence class thus has two degrees of freedom. This metric can be expressed as a percentage to convey the amount of underdetermination relative to the number of variables in the system. Again, for the equivalence class in Figure 2.3, there are $2/(2^{\binom{3}{2}}) \approx 33\%$ of the degrees of freedom remaining. Once enough

² The blank space between the two variables is intentional; it is meant to call attention to the fact that the corresponding nodes in the graph lack any type of edge between them.

constraints have been supplied to prune an equivalence class to only one graph, zero degrees of freedom remain. This pruning of the equivalence class thus provides an analytic expression for Popper’s conception of science based on falsifiability [Pop59].

Given a set of (in)dependence relations (Chapters 4) expressed as constraints on causal structure, we can use the causal discovery technique discussed in Chapter 5 to obtain the degrees of freedom for the equivalence class that is consistent with the constraints. For the case where we assume that the true causal graph is a DAG, the approach is given in Algorithm 1 and proceeds as follows. We define the set \mathbf{K} as the set of causal-structure constraints obtained for a system with the set of variables \mathbf{V} . For each pair of variables $\{X, Y\}$ in the system, we run Clingo once for every degree of freedom that can exist between X and Y : in a given run, we input the constraints in \mathbf{K} as well as one additional constraint, which encodes the particular degree of freedom being tested. The degrees of freedom $X \rightarrow Y$, $X \leftarrow Y$, and $X \perp\!\!\!\perp Y$ are encoded by the sets of ASP constraints $\{\text{edge}(X, Y) .\}$, $\{\text{edge}(Y, X) .\}$, and $\{-\text{edge}(X, Y) ., -\text{edge}(Y, X) .\}$, respectively. The hyphens (-) in the last set indicate negation. In each run, Clingo returns either SATISFIABLE or UNSATISFIABLE, indicating whether the potential degree of freedom occurs at least once in the equivalence class—i.e., whether the edge relation exists in any of the consistent causal graphs. A system with N variables and three possible relations between each pair of variables will require $3 \binom{N}{2}$ runs of Clingo to fully determine the degrees of freedom. Therefore, this procedure splits the set of all possible edge relations into two sets: (1) the degrees of freedom, each of which appears in at least one graph in the equivalence class, and (2) the relations that have been ruled out by the constraints. This procedure can be extended to consider cyclic causal graphs by including the degree of freedom indicated by the constraint set $\{\text{edge}(X, Y) ., \text{edge}(Y, X) .\}$.

In the next chapter, I discuss how causal graphs can help to plan experiments. With an equivalence class, this task can be framed as the search for an experiment that could minimize causal underdetermination. As an aid to causal reasoning, the degrees of freedom allow one to make fairly strong inferences regarding *every* graph in the equivalence class—even if it would be too computationally expensive to compute every graph explicitly. By giving scientists a more “global” perspective on the range of causal explanations that remain viable, such inferences can help to identify when a potential experiment would be either interesting or uninformative. This approach to experiment planning

Data: \mathbf{K} : set of causal-structure constraints over the set of variables \mathbf{V}

Result: \mathbf{D} : set of degrees of freedom for each pair of variables in the equivalence class

$\mathbf{D} \leftarrow \emptyset$;

for each pair of variables $\{X, Y\} \in \mathbf{V}$ **do**

for each set of constraints, \mathbf{K}_d , encoding a potential degree of freedom for $\{X, Y\}$ **do**

$s \leftarrow$ satisfiability of constraint set $(\mathbf{K} \cup \mathbf{K}_d)$;

if $s = \text{SATISFIABLE}$ **then**

$\mathbf{D} \leftarrow (\mathbf{D} \cup \mathbf{K}_d)$;

end

end

end

Algorithm 1: Deriving the degrees of freedom for an equivalence class

allows one to categorize hypotheses with respect to the range of causal interpretations that remain viable, given the evidence. These categories, discussed in § 7.2, have enormous consequences for experiment planning. In § 8.2, I demonstrate how degree-of-freedom analyses were used to identify conflicts in the ResearchMaps database. I also discuss how this approach can be used to draw inferences regarding phenomena that did not appear together in any one study; instead, such inferences arise out of the synthesis of multiple studies.

CHAPTER 7

Selecting the next experiment

When scientists plan their next experiment, they can view the activity as trying to either maximize evidence or minimize uncertainty. This chapter shows how these two complementary approaches can be guided by research maps and causal graphs.

7.1 Maximizing evidence in research maps

As a representation of empirical evidence, a research map provides guidelines for selecting experiments that will maximize evidence. The Bayesian model used to compute a research map's CEIs (§ 6.1) can be queried to obtain the particular study design(s) that would most effectively add evidence to a particular edge or pathway. Thus, studies can be ranked by the value of the evidence that they could potentially yield [MWW17a].

Given that convergence and consistency are used to gauge evidence in research maps, these principles can also be used to determine which studies could most effectively strengthen or weaken the evidence for a particular edge. For example, if the evidence for an edge is based solely on a positive intervention, the principle of convergence would suggest that negative interventions and non-intervention studies could be used to strengthen the evidence for that edge. Additionally, the principle of consistency would suggest that repetitions of any one of these studies could strengthen the evidence. This reasoning represents a straightforward approach commonly used by scientists to plan their research. Beyond just a single edge, these integration rules can be extended to an entire research map. To facilitate the presentation of these principles, I limit the discussion below to research maps that contain only three nodes, representing part of a signaling pathway or biological cascade.

It is important to remember that studies are usually carried out with reference to a specific

hypothesis that is commonly suggested by findings and theories. In research maps, hypotheses are represented by hypothetical edges. Unlike edges representing empirical studies, hypothetical edges have no CEI or study symbols (Figure 2.1). Hypothetical edges can thus organize and structure empirical edges that are based on actual studies. Although the causal relations represented by hypothetical edges cannot always be directly tested—we may lack the required tools—they nevertheless inform the choice among feasible studies by contextualizing empirical results within specific theories or interpretations.

With a given research map, we can use a number of principles, including the *pioneering* rule, to develop its evidence. This pioneering rule states that when a research map's edges imply the existence of an edge that spans other edges, testing this edge can significantly inform the model. For example, if we have a research map with empirical edges $X \rightarrow Y \rightarrow Z$, then designing a study to test the edge $X \rightarrow Z$ will likely be instructive as to whether X contributes to Z . Finding that interventions on X reliably affect Z , for example, will provide further evidence for the existence of a pathway from X to Z .

Having considered all the pairwise edges in a research map, we then refer to what is called the *weakest-link* rule. This rule simply states that edges with the lowest CEI should receive the most attention when designing studies to assess a given research map. Using the example above, if the $X \rightarrow Y$ edge has a CEI of 0.250 while the $Y \rightarrow Z$ edge has a CEI of 0.125, the weakest-link rule states that we should further test the $Y \rightarrow Z$ edge first. Note that once a particular edge has been selected for additional studies, the single-edge integration rules of convergence and consistency (§ 6.1) provide guidelines for selecting the optimal study to perform.

There are cases when the above rules cannot identify a single study that is optimal: there may be two or more study classes (e.g., positive and negative interventions) that could (potentially) provide equally convergent and consistent evidence, given the studies that have already been performed. In such cases we refer to the rule of *multi-edge convergence*. This rule states that when given a choice between (potentially) equally convergent study classes, we should select the class that is least represented among studies recorded for the entire research map. The rationale for this rule is that increasing the methodological diversity of a set of findings will lower the chances of systematic artifacts. For example, the prevalence of negative interventions depicted in Figure 2.1 would en-

courage the use of positive interventions, as well as non-interventions, to study this system further.

These rules—(single-edge) convergence and consistency, the pioneering rule, the weakest-link rule, and multi-edge convergence—provide guidelines for experiment planning when working with research maps. These rules attempt to make explicit and quantitative the epistemological strategies commonly used by biologists. Just as Hill’s criteria for causation [Hil65] provided qualitative descriptions of causality that were later formalized in graphical models, these experiment-selection criteria codify commonsense intuitions that currently guide experiment planning in biology.

7.2 Minimizing underdetermination in causal graphs

When we interpret an experiment’s result, it is prudent to be conservative: we should draw only the most likely conclusions, or else we risk exaggerating the result’s impact. But when we plan the next experiment, it is valuable to be aggressive: we should target the part of the system about which there is the most ambiguity, the most conflicting information. We learn the most from experiments whose results we cannot predict—or whose results we predicted incorrectly—than from experiments whose results we can predict confidently [NKS16].

As a complement to selecting the experiment that could maximize evidence (§ 7.1), one can try to identify the experiment that could minimize uncertainty. An equivalence class of causal graphs—the set of causal explanations that remain consistent with what is known—provides an analytic basis for posing this task of minimizing uncertainty. In the context of causal graphs, this uncertainty is more accurately referred to as *underdetermination* [May13].

An equivalence class contains every causal explanation that remains consistent with what is known about the system. For instance, in a particular equivalence class, some graphs might have the edge $X \rightarrow Y$, others $X \leftarrow Y$. This diversity of causal structures indicates the location and magnitude of underdetermination about a system’s causal structure, and thus which evidence about the system is lacking. When quantified, this underdetermination provides an analytic basis for selecting the next experiment.

Given a knowledgebase of constraints on causal structure, the pipeline in Figure 1.4 provides a way to place a given hypothesis in one of three categories, with crucial distinctions:

1. *The hypothesis is consistent with **none** of the causal graphs in the equivalence class.* This kind of hypothesis should be pursued only if we are confident that one or more constraints in the current knowledgebase are incorrect. The hypothesis is then useful insofar as it identifies which constraints in the knowledgebase could be refuted. Otherwise, given the current knowledgebase, we would fail to find even one causal graph that is consistent with this kind of hypothesis.
2. *The hypothesis is consistent with **all** the causal graphs in the equivalence class.* Although this kind of hypothesis produces accurate predictions about the system, it is equally unhelpful as the first kind with respect to experiment selection: this hypothesis should not be tested empirically unless we believe there to be a flaw in our current knowledgebase and wish to refute one or more of its constraints. The reason is that if a hypothesis is consistent with *all* the causal graphs in the equivalence class, it already follows logically from the knowledgebase; the logical proposition that expresses the hypothesis is thus true for all solutions (i.e., causal graphs). In propositional logic, it is said to be in the *backbone* of the satisfying formula [HHE13].
3. *The hypothesis is consistent with **some** (not all) of the causal graphs in the equivalence class.* This kind of hypothesis is most worth pursuing empirically. The experiment’s result—which the current knowledgebase cannot predict with certainty—is guaranteed to prune the equivalence class, bringing us closer to the true causal graph.

We can easily test which category a hypothesis belongs to. First, we express the hypothesis as a hypothetical edge in a research map, which is translated into a formal causal-structure constraint. Second, we compute whether the conjunction of this hypothetical constraint and the rest of the knowledgebase is logically satisfiable—that is, whether the conjunction is consistent with none, all, or some of the causal graphs in the equivalence class. As with the degree-of-freedom analysis, this procedure does not require the SAT solver to perform the expensive computation of enumerating every graph in the equivalence class. Instead, we can simply compute whether the hypothesized constraint is satisfiable, as a binary condition. If the answer is no, then we know that the hypothesis falls into the first category: it is consistent with none of the causal graphs in the equivalence class. If the answer is yes, then we must distinguish between whether the hypothesis is consistent with some or all of the graphs. We do this by querying for the satisfiability of the hypothesis’s negation. If

the hypothesis's negation *cannot* be satisfied by any of the graphs, then we know that the hypothesis falls into the second category: it is consistent with all the causal graphs in the equivalence class. If the negation *can* be satisfied by at least one graph, then we know that the hypothesis falls into the third category: it is consistent with some (not all) of the causal graphs in the equivalence class. Therefore, any hypothesis, expressed as a causal-structure constraint, can be categorized with only one or two queries to a SAT solver. Despite the enormous consequences that this categorization has on experiment planning, it is usually infeasible for a scientist to manually compute which category a hypothesis belongs to.

It can also be instructive to analyze the particular pattern of degrees of freedom that exists for each pair of variables in the equivalence class. Table 7.1 lists each possible pattern and a general interpretation of what each pattern suggests about the causal interpretations that remain viable. After performing the degree-of-freedom analysis described in § 6.2, a scientist can inspect the viable degree-of-freedom patterns to see precisely which edge relations have been ruled out by the available evidence. If, based on other background knowledge or assumptions, it is believed that additional edges could be ruled out, the (in)dependence relation(s) that would be required to further prune the equivalence class can be identified. Experiments can then be planned to test these hypothesized (in)dependence relations.

Based on the degrees of freedom that remain viable, some experiments will be more effective than others in their ability to further prune the equivalence class. For instance, if two variables X and Y exhibit pattern #5 in Table 7.1, we can conclude that if a third variable Z is a *necessary* mediator between X and Y , this variable does not appear in the set of variables for which the degree-of-freedom analysis was performed: every causal graph in the equivalence has a direct edge between X and Y . Note that this kind of analysis allows us to make fairly strong statements about the entire equivalence class without having to enumerate every one of its causal graphs explicitly. Based on this kind of analysis, one can thus plan the next experiment with a better sense of the causal explanations that remain viable, given the current evidence. For each degree-of-freedom pattern that can exist for a pair of variables, $\{X, Y\}$, Table 7.2 lists the suggested intervention sets that will most effectively distinguish between the pair's remaining degrees of freedom.

The degrees of freedom can be used as the basis for experiment-selection methods. Below, I

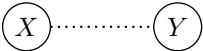
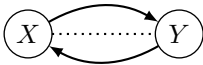
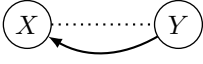
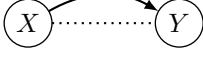



Degree-of-freedom pattern	Interpretation
1. 	not connected
2. 	not connected, or connected in either direction
3. 	not connected, or connected in one direction
4. 	not connected, or connected in one direction
5. 	connected in either direction
6. 	connected in one direction
7. 	connected in one direction

Table 7.1: The possible patterns for degrees of freedom in an equivalence class over DAGs. Each pattern is associated with an interpretation given in plain language. In this table, “connected” implies “*directly* connected”: note that all these patterns—including the first—still allow for an *indirect* path between X and Y via other nodes in the graph. A scientist can inspect these patterns to see which edge relations have been ruled out by the available evidence. This information can motivate the selection of additional experiments (Table 7.2).

present two methods: the first is based primarily on the degrees of freedom of the equivalence class; the second is based on the degrees of freedom and also an expectation metric. The first method is computationally less expensive because it does not require the enumeration of every causal graph in the equivalence class, as discussed in § 6.2. The second method requires more computation, but its suggestions are correspondingly more informed, leading to more efficient causal discovery. Both methods can be performed with input from a domain expert: after each empirical result is added to the knowledgebase, the suggestions based on these graphical metrics should be evaluated with respect to the full diversity of constraints on experiment planning that currently only a human being can consider.

Algorithm 2 gives an experiment-selection method based on the degrees of freedom. First, for each pair of variables in the system, $\{X, Y\}$, we obtain $n_{X,Y}$, the number of degrees of freedom in the equivalence class \mathbf{E} for the pair $\{X, Y\}$. Next, for the $(X, Y, n_{X,Y})$ three-tuple with the largest $n_{X,Y}$, we randomly choose one of the suggested experiments for the pair’s degrees of freedom, $\mathbf{D}_{X,Y}$, as given in Table 7.2. (If multiple three-tuples have the same maximum $n_{X,Y}$, we choose one randomly.) The experiments in Table 7.2 are chosen to be maximally informative, given the degrees of freedom that remain viable. For example, if the relations $X \rightarrow Y$ and $X \perp\!\!\!\perp Y$ are the remaining degrees of freedom, we do not suggest an intervention on Y , as this would remove the $X \rightarrow Y$ edge, rendering the two relations indistinguishable [Pea09]. Because this algorithm suggests an experiment given a set of experiments that have already been performed, additional bookkeeping is done to ensure that experiments are not repeated unnecessarily [RFF15] (see the while-loop).

When it is possible to explicitly compute every causal graph in the equivalence class, we can improve on the efficiency of Algorithm 2: Algorithm 3 gives an experiment-selection method that incorporates an expectation metric.¹ As with Algorithm 2, this method uses the degrees of freedom of the equivalence class. But here the intuition is also grounded in expectation maximization. First, for each pair of variables in the system, $\{X, Y\}$, and for each possible degree of freedom, d , we obtain $m_{X,Y}^d$, the number of graphs in the equivalence class \mathbf{E} that assign the degree of freedom d to the pair $\{X, Y\}$. We use this quantity to calculate the empirical probability of a graph in the equivalence class having that particular degree of freedom: $\frac{m_{X,Y}^d}{|\mathbf{E}|}$. We also calculate the number of graphs that would be eliminated from the equivalence class if we were to learn that this degree of freedom was the actual relation taken by that pair of variables in the true causal graph: $|\mathbf{E}| - m_{X,Y}^d$. This empirical probability, $\frac{m_{X,Y}^d}{|\mathbf{E}|}$, is multiplied by its associated “reward,” $|\mathbf{E}| - m_{X,Y}^d$, yielding the pair’s expectation for a given d : $e_{X,Y}^d = \frac{m_{X,Y}^d}{|\mathbf{E}|} (|\mathbf{E}| - m_{X,Y}^d)$. Next, for the $(X, Y, d, e_{X,Y}^d)$ four-tuple with the highest expectation, we randomly choose one of the suggested experiments for d , as given in the last three rows of Table 7.2. (If multiple four-tuples have the same maximum $e_{X,Y}^d$, we choose one randomly.) As is true for Algorithm 2, additional bookkeeping is performed to ensure that experiments are not repeated unnecessarily [RFF15].

Although these experiment-selection heuristics will not achieve known limits of efficiency

¹ I am grateful to Justin Wood for suggesting this approach.

Data: \mathbf{K} : set of causal-structure constraints over the set of variables \mathbf{V} ;

\mathbf{P} : set of experiments performed to obtain \mathbf{K}

Result: s : experiment suggested on the basis of \mathbf{K} and \mathbf{P}

$\mathbf{E} \leftarrow$ equivalence class (maximally) consistent with \mathbf{K} (Chapter 5);

$\mathbf{D} \leftarrow$ degrees of freedom for each pair of variables in \mathbf{E} (Algorithm 1);

$\mathbf{R} \leftarrow \emptyset$;

for each pair $\{X, Y\} \in \mathbf{V}$ **do**

$n_{X,Y} \leftarrow$ number of degrees of freedom in \mathbf{E} for $\{X, Y\}$;

$\mathbf{R} \leftarrow \mathbf{R} \cup \{(X, Y, n_{X,Y})\}$;

end

rank \mathbf{R} by $n_{X,Y}$ in descending order;

$c \leftarrow 0$;

while $c < \max | \mathbf{S}_{\mathbf{D}_{X,Y}} |$ **do**

for each $(X, Y, n_{X,Y}) \in \mathbf{R}$ **do**

$\mathbf{S}_{\mathbf{D}_{X,Y}} \leftarrow$ set of experiments suggested according to $\mathbf{D}_{X,Y}$ (Table 7.2);

if $| \mathbf{S}_{\mathbf{D}_{X,Y}} \cap \mathbf{P} | \leq c$ **and** $| \mathbf{S}_{\mathbf{D}_{X,Y}} \cap \mathbf{P} | < | \mathbf{S}_{\mathbf{D}_{X,Y}} |$ **then**

$s \leftarrow s \in (\mathbf{S}_{\mathbf{D}_{X,Y}} - \mathbf{P})$;

 return s ;

end

end

$c \leftarrow c + 1$

end

return random experiment from set of possible experiments not in \mathbf{P} ;

Algorithm 2: Experiment selection based on degrees of freedom

for experiment selection and causal discovery, they are grounded in the graphical representations that scientists already use to express causal mechanisms. As a result, scientists can readily interpret the algorithm's rationale for suggested experiments in the context of the graphical models that they consider to be viable. Although any experiment, if executed properly, can yield useful information regarding a system, strategic experiment selection—even if guided simply by heuristics—can

Data: \mathbf{K} : set of causal-structure constraints over the set of variables \mathbf{V} ;

\mathbf{P} : set of experiments performed to obtain \mathbf{K}

Result: s : experiment suggested on the basis of \mathbf{K} and \mathbf{P}

$\mathbf{E} \leftarrow$ equivalence class (maximally) consistent with \mathbf{K} (Chapter 5);

$\mathbf{D} \leftarrow$ degrees of freedom for each pair of variables in \mathbf{E} (Algorithm 1);

$\mathbf{R} \leftarrow \emptyset$;

for each pair $\{X, Y\} \in \mathbf{V}$ **do**

for each degree of freedom $d \in \mathbf{D}_{X,Y}$ **do**

$m_{X,Y}^d \leftarrow$ number of graphs in \mathbf{E} with degree of freedom d for X, Y ;

$e_{X,Y}^d \leftarrow \frac{m_{X,Y}^d}{|\mathbf{E}|} (|\mathbf{E}| - m_{X,Y}^d)$;

$\mathbf{R} \leftarrow \mathbf{R} \cup \{(X, Y, d, e_{X,Y}^d)\}$;

end

end

rank \mathbf{R} by $e_{X,Y}^d$ in descending order;

$c \leftarrow 0$;

while true do

for each $(X, Y, d, e_{X,Y}^d) \in \mathbf{R}$ **do**

$\mathbf{S}_{D_{X,Y}} \leftarrow$ set of experiments suggested according to d (Table 7.2);

if $|\mathbf{S}_{D_{X,Y}} \cap \mathbf{P}| \leq c$ **and** $|\mathbf{S}_{D_{X,Y}} \cap \mathbf{P}| < |\mathbf{S}_{D_{X,Y}}|$ **then**

$s \leftarrow s \in (\mathbf{S}_{D_{X,Y}} - \mathbf{P})$;

 return s ;

end

end

$c \leftarrow c + 1$

end

Algorithm 3: Experiment selection based on degrees of freedom and expectation

save considerable amounts of work toward identifying a system's true causal graph. The simulations in the next chapter quantify these savings by comparing the experiment-selection policies of Algorithms 2 and 3 to random experiment selection.


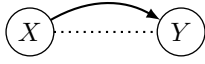
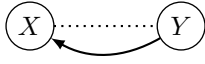


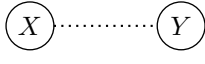

Degree-of-freedom pattern, $D_{X,Y}$	Suggested experiments, $S_{D_{X,Y}}$	$S_{D_{X,Y}}$ in research maps
	$J = \emptyset$ $J = \{X\}$ $J = \{Y\}$	$\emptyset^\uparrow, \emptyset^\downarrow, \uparrow X, \downarrow X, \uparrow Y, \downarrow Y$
	$J = \emptyset$ $J = \{X\}$	$\emptyset^\uparrow, \emptyset^\downarrow, \uparrow X, \downarrow X$
	$J = \emptyset$ $J = \{Y\}$	$\emptyset^\uparrow, \emptyset^\downarrow, \uparrow Y, \downarrow Y$
	$J = \{X\}$ $J = \{Y\}$	$\uparrow X, \downarrow X, \uparrow Y, \downarrow Y$
	$J = \{X\}$	$\uparrow X, \downarrow X$
	$J = \{\emptyset\}$	$\emptyset^\uparrow, \emptyset^\downarrow$
	$J = \{Y\}$	$\uparrow Y, \downarrow Y$

Table 7.2: The experiments that would be most informative with respect to a pair of variables, given their particular degree-of-freedom pattern in an equivalence class. These suggested experiments inform the experiment-selection methods given in Algorithms 2 and 3. In the third column, the variable that appears after each intervention symbol (\uparrow or \downarrow) is the agent that is intervened on.

CHAPTER 8

Evaluation

8.1 Reasoning with structural patterns in research maps

Interpreting and reporting scientific results usually involve some abstraction away from what was actually done in the laboratory. For example, in neuroscience studies that address how the protein CREB affects memory, the phenomenon of memory is not measured directly, as if, like temperature, it can be detected physically by a sensor. Instead, researchers measure another variable, such as the time required to complete a maze, with the interpretation that this measurement is a proxy for memory.

If scientists were restricted to reporting only what literally happened in the laboratory, articles would often fail to articulate how the collected data fit into an emerging theory of the system. But when scientists summarize their findings in the literature, it is certainly possible for them to interpret and abstract excessively. In the example above, a more conservative interpretation might be that a protein affects *spatial* memory in particular, while it may not have an effect on other forms of memory, such as fear memory. This idea is illustrated by high-density lipoprotein (HDL) and low-density lipoprotein (LDL): although they are both forms of cholesterol, they have opposite effects on heart disease. Therefore, if researchers study the effect of cholesterol on heart disease without differentiating between HDL and LDL, they will likely obtain conflicting results, even if each individual study is well executed [SS04].

It would be very useful to analytically detect when causal variables have been misspecified, as is true when scientists excessively rely on interpretation or abstraction [Ebe16]. This problem can be addressed by annotating research articles using the research-map schema and looking for structural conflicts in the resulting map: such conflicts signal logical inconsistencies in the story that is being presented regarding the data. Of course, some of these conflicts will arise due to poorly executed

experiments that yield erroneous (and thus conflicting) results. But many conflicts can be resolved by revising the research map to more faithfully represent what literally happened in the course of the study—that is, by interpreting less, and by specifying the causal variable with more precision.

Consider the following example of how research maps can be used to identify and correct misspecified causal variables. Figure 8.1 shows a particular pattern of research-map edges that was found to permit a particular inference. Given the research-map edges $A \rightarrow B$, $A \cdots \bullet C$, $D \rightarrow C$, and $D \cdots \bullet B$, it follows that there can be neither an excitatory nor inhibitory path between A and D , whether direct or indirect. For instance, an $A \rightarrow D$ edge would create the path $A \rightarrow D \rightarrow C$, creating a correlation between A and C ; however, this would conflict with the edge $A \cdots \bullet C$, which states that A and C are independent. A symmetric argument applies for the other disallowed edge: a $D \rightarrow A$ edge would create the path $D \rightarrow A \rightarrow B$, creating a correlation between D and B ; however, this would conflict with the edge $D \cdots \bullet B$, which states that D and B are independent. Figure 8.2 shows the Neo4j query that can be used to find a violation of this inference in the ResearchMaps database. This query was used to identify conflicts in the neuroscientist Alcino J. Silva’s research map for the article “A temporal shift in the circuits mediating retrieval of fear memory,” published in *Nature* [DQQ15] (Figure 8.3). Dr. Silva then revised the research map to resolve the conflicts, producing the research map in Figure 8.4. The two research maps contain the same number of experiments, so they both faithfully represent the work performed in the laboratory; however, they instantiate different nodes and thus imply different causal explanations for the empirical results. This revision required the When properties of some nodes to be specified with greater precision, thereby avoiding excessive abstraction—a version of the LDL–HDL problem described above.

We note two important advantages of this approach to conflict detection. The first is that these types of analyses can be “recycled”: once an inference has been made, a database of research maps can be queried for other instances of the structural template—the specific configuration of nodes and edges—that permitted the original inference; note that the inference will hold regardless of the identity of the nodes involved. The second advantage is that combining structural information from multiple articles allows one to make inferences about biological phenomena that may have never even appeared together in the same experiment, or which were never discussed together in a single article [Dan05]. Although conflicts among results may be apparent if they occur in a single

article, a conflict’s structural components may be spread out across many articles, making it difficult to find. It is thus extremely difficult to anticipate where such conflicts might arise, and it is challenging to notice ones derived from the synthesis of many articles, unless one is already looking for a specific pattern—particularly as the patterns become increasingly complex. In fact, another instance of the pattern in Figure 8.1 was found in the ResearchMaps database; however, in this case the pattern arose only when the research maps of two separate articles were merged, as described in § 8.2.

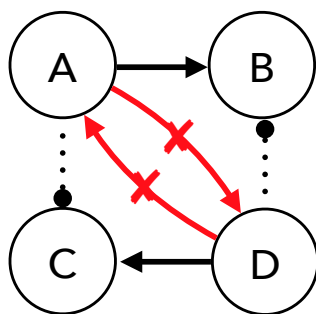


Figure 8.1: A pattern of research-map edges that imply a lack of paths between the variables A and D . The Neo4j query for detecting a conflict with this pattern is given in Figure 8.2. Exactly this conflict was found in the original research map for a neuroscience article [DQQ15], as highlighted in Figure 8.4.

8.2 Reasoning with degrees of freedom in causal graphs

We used the meta-analytic pipeline shown in Figure 1.4 to synthesize the evidence in two neuroscience articles [GFF98, THT96]. The phenomena addressed in these articles were to become the nodes in a causal graph; we thus chose articles with partially overlapping sets of phenomena, simulating the analogous situation of causal discovery with partially overlapping datasets [HEJ14]. Alcino Silva annotated these two articles, yielding a research map for each (Figures 8.5 and 8.6). Because the ASP-based procedure described in Chapter 5 does not scale well past eight variables, subsets of the results involving seven variables were extracted and merged into a single research map for further analysis (Figure 8.7). This merged research map allowed for the demonstration of meta-analytic causal discovery in a small but nonetheless real biological system.

```

MATCH (w:Experiment)-[:agent]->(a:NeuroIaxTerm)<-[:agent]-(x:Experiment)-[:
  target]->(c:NeuroIaxTerm)<-[:target]-(y:Experiment)-[:agent]->(d:
  NeuroIaxTerm)<-[:agent]-(z:Experiment)-[:target]->(b:NeuroIaxTerm)<-[:
  target]-(w:Experiment),
(a)<-[:agent]-(r:Experiment)-[:target]->(e:NeuroIaxTerm),
(e)<-[:agent]-(q:Experiment)-[:target]->(d:NeuroIaxTerm)
WHERE (w.conclusion='No Relation')
AND NOT (x.conclusion='No Relation')
AND (y.conclusion='No Relation')
AND NOT (z.conclusion='No Relation')
AND NOT (r.conclusion='No Relation')
AND NOT (q.conclusion='No Relation')
AND ID(a)<ID(d)
RETURN a.What,a.Where,a.When,e.What,e.Where,e.When,d.What,d.Where,d.When;

```

Figure 8.2: The Neo4j query used to find the conflicts that are highlighted in Figure 8.1, and which appeared in the original version of a research map (Figure 8.3).

The two research maps were translated into formal causal-structure constraints as described in Chapter 4, yielding ten constraints over the seven variables (Table 8.1). These constraints were then used to identify consistent causal explanations, as described in Chapter 5. Because there are over *one billion* possible causal graphs for seven variables [Rob73], we did not enumerate every graph in the equivalence class, which would have been expensive to compute, and impractical for a domain expert to assess. Instead, we performed a degree-of-freedom analysis (§ 6.2), which characterizes the equivalence class in a way that not only is faster to compute but also yields a digestible and actionable visualization. Our degree-of-freedom analysis thus showed which edge relations remain viable for this system: the annotated results are consistent with 42 (66%) degrees of freedom, and inconsistent with 21 (33%). Each degree of freedom took on the order of 30 seconds to compute using a 2.4 GHz Intel Core i5 processor. Figure 8.8 shows in black the edge relations that remain viable; in red are edge relations that have been ruled out. Note that Figure 8.9—one of the con-

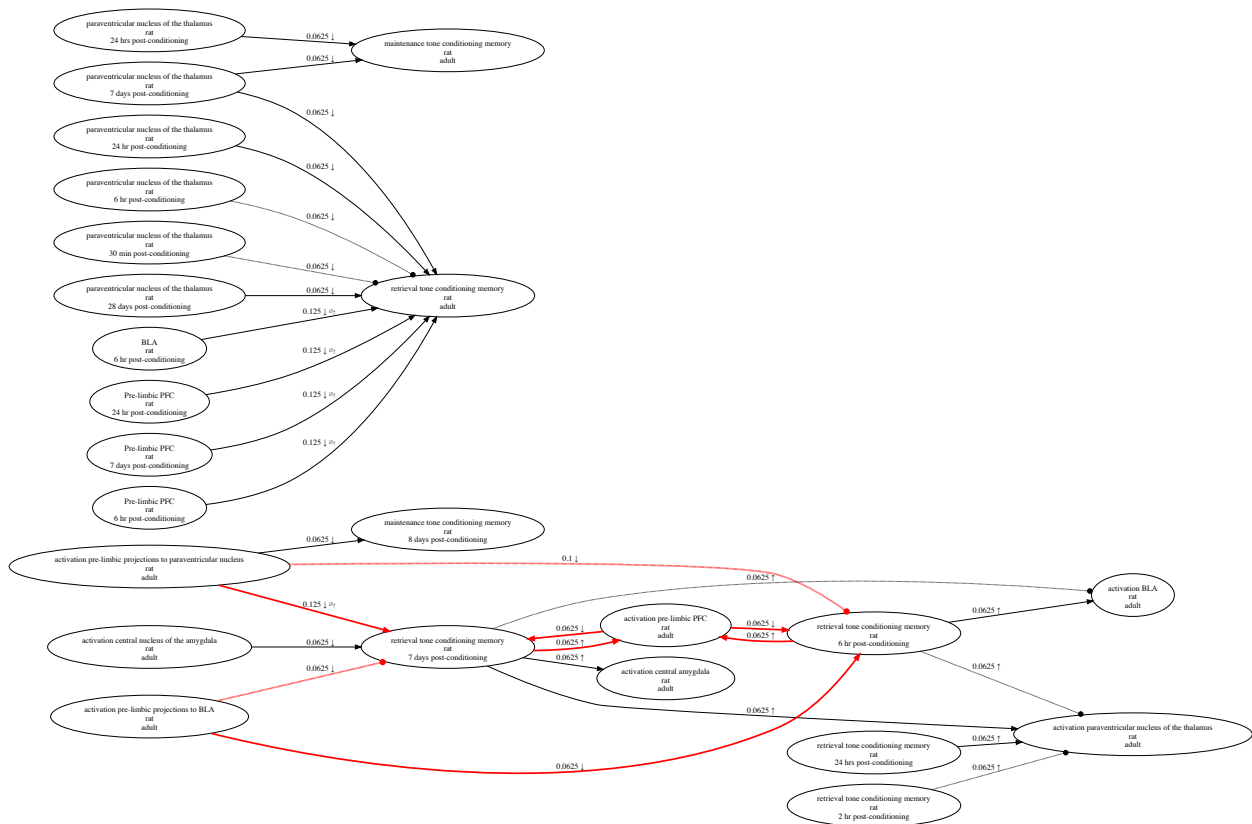


Figure 8.3: Alcino Silva’s original research map for Do-Monte et al. [DQQ15], which contained conflicts. This research map was revised to resolve its conflicts, yielding the research map in Figure 8.4.

sistent causal graphs—contains a particular set of the viable degrees of freedom from Figure 8.8. Although the model space for causal graphs is very large, this use case showed that a set of just ten constraints is sufficient to eliminate a third of the degrees of freedom from a system with over one billion possible graphs. A domain expert can inspect the remaining degrees of freedom to determine which constraints—and thus which experiments—are needed to eliminate additional edges from consideration.

Even though this analysis does not enumerate every causal graph in the equivalence class, the specific degree-of-freedom patterns that exist between each pair of nodes allow one to make strong statements regarding *all* the graphs in the equivalence class. Consider the CaMKII and NMDAR1 nodes in Figure 8.8; the degrees of freedom between these nodes have pattern #1 in Table 7.1: the two directed edges are ruled out (red), and the dotted line—indicating the absence of an edge in

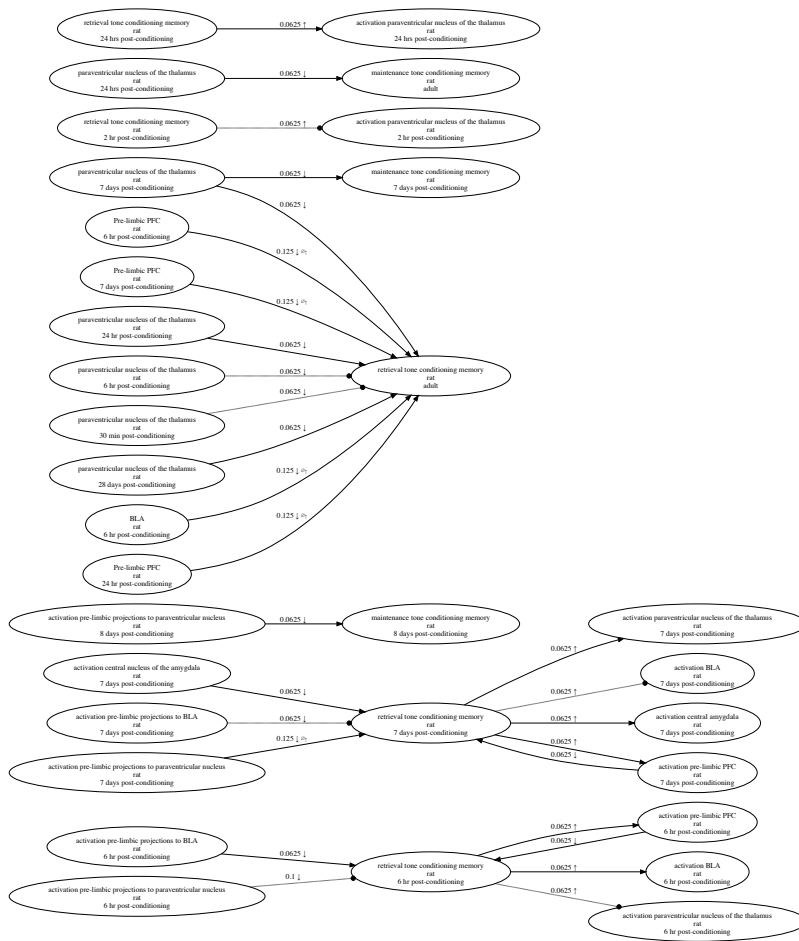


Figure 8.4: A revised version of Alcino Silva’s original research map for Do-Monte et al. [DQQ15] (Figure 8.3) that resolves conflicts present in the original version.

the causal graph—remains viable (black). This pattern implies that these two nodes cannot have a direct path connecting them in the true causal graph. In fact, because the constraints involving these nodes had empty pre conditioning sets, we can state further that these two nodes also cannot have an indirect path connecting them.¹

The research-map edges that gave rise to this degree-of-freedom pattern is exactly the pattern in Figure 8.1, discussed in § 8.1. Whereas this pattern was used to detect a conflict in a single research map, it was also used to make an inference about causal structure—and thus to predict the outcome of experiments—in a set of results that span multiple articles (in this case, two). Specifically, this pattern of research-map edges (and degrees of freedom) imply that any study involving

¹ I am grateful to Frederick Eberhardt for pointing out this inference. If there are any errors in the discussion of this finding, they are mine.

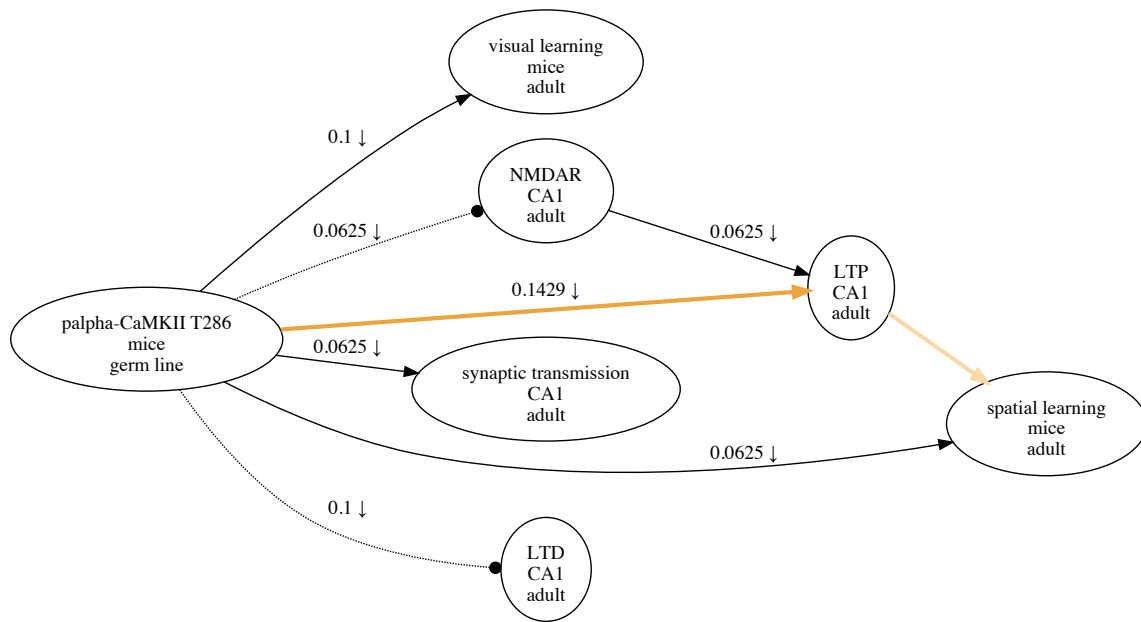


Figure 8.5: Alcino Silva's research map for Giese et al. [GFF98].

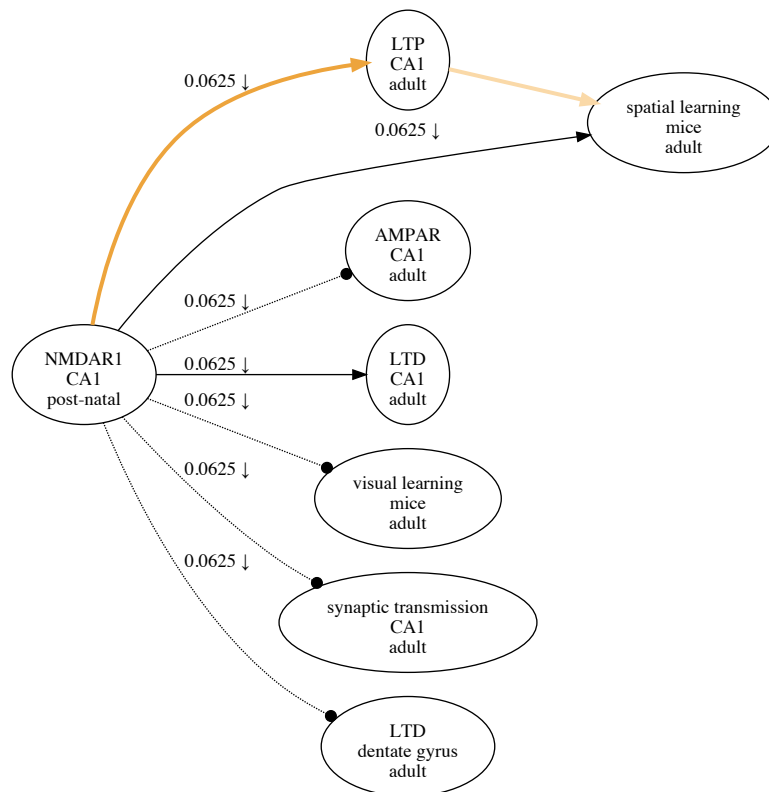


Figure 8.6: Alcino Silva's research map for Tsien et al. [THT96].

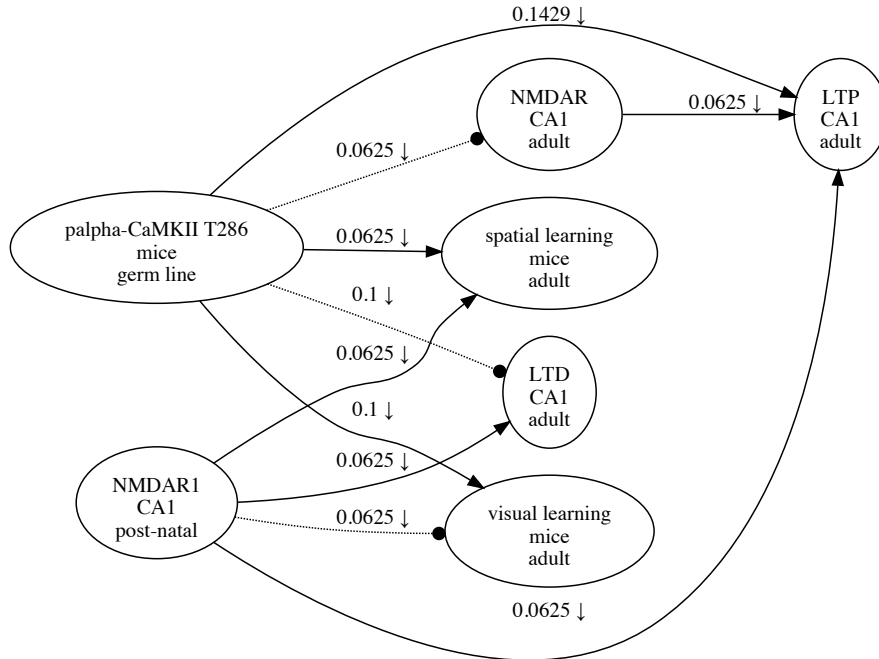


Figure 8.7: A research map showing a subset of the results in Figure 8.5 and Figure 8.6 merged into a single research map.

these two variables would lead to an independence relation. This hypothesis was deemed plausible by a senior neuroscientist (Alcino Silva). As with the analysis described in § 8.1, the degree-of-freedom pattern on which this inference is based can be “recycled”; the same inference will apply to other pairs of nodes that share this degree-of-freedom pattern.

8.3 Simulations of experiment-selection algorithms

The experiment-selection policies given in Algorithms 2 and 3 were evaluated using the following simulation. First, one of the 543 possible DAGs over four variables was set as the true graph. Before any experiments were simulated, the equivalence class trivially contained every possible graph. To simulate how researchers learn about a system through repeated experimentation, we sampled experimental designs according to three different policies: at each iteration, we chose the next experiment (1) randomly, (2) according to Algorithm 2, and (3) according to Algorithm 3. The correct result of each experiment was returned by an oracle that assumed causal sufficiency and had access to the true causal graph. Each experiment’s result was added to a growing list of constraints,

(in)dependence relation	ASP constraint
$p\alpha\text{-CaMKII T286} \perp\!\!\!\perp \text{LTD} \mid \emptyset \parallel p\alpha\text{-CaMKII T286}$	$\text{indep}(1, 2, 0, 1, 124, 1)$.
$p\alpha\text{-CaMKII T286} \not\perp\!\!\!\perp \text{spatial learning} \mid \emptyset \parallel p\alpha\text{-CaMKII T286}$	$\text{dep}(1, 3, 0, 1, 122, 1)$.
$p\alpha\text{-CaMKII T286} \not\perp\!\!\!\perp \text{LTP} \mid \emptyset \parallel p\alpha\text{-CaMKII T286}$	$\text{dep}(1, 4, 0, 1, 118, 1)$.
$\text{LTP} \not\perp\!\!\!\perp \text{NMDAR} \mid \emptyset \parallel \text{NMDAR}$	$\text{dep}(4, 5, 0, 16, 103, 1)$.
$p\alpha\text{-CaMKII T286} \perp\!\!\!\perp \text{NMDAR} \mid \emptyset \parallel p\alpha\text{-CaMKII T286}$	$\text{indep}(1, 5, 0, 1, 110, 1)$.
$p\alpha\text{-CaMKII T286} \not\perp\!\!\!\perp \text{visual learning} \mid \emptyset \parallel p\alpha\text{-CaMKII T286}$	$\text{dep}(1, 6, 0, 1, 94, 1)$.
$\text{visual learning} \perp\!\!\!\perp \text{NMDAR1} \mid \emptyset \parallel \text{NMDAR1}$	$\text{indep}(6, 7, 0, 64, 31, 1)$.
$\text{LTD} \not\perp\!\!\!\perp \text{NMDAR1} \mid \emptyset \parallel \text{NMDAR1}$	$\text{dep}(2, 7, 0, 64, 61, 1)$.
$\text{spatial learning} \not\perp\!\!\!\perp \text{NMDAR1} \mid \emptyset \parallel \text{NMDAR1}$	$\text{dep}(3, 7, 0, 64, 59, 1)$.
$\text{LTP} \not\perp\!\!\!\perp \text{NMDAR1} \mid \emptyset \parallel \text{NMDAR1}$	$\text{dep}(4, 7, 0, 64, 55, 1)$.

Table 8.1: The (in)dependence relations derived from the research map in Figure 8.7, along with their ASP encodings. For brevity, only the What property of each variable is listed. The following integer indices are used in the first two arguments of each ASP constraint: 1: $p\alpha\text{-CaMKII T286}$; 2: LTD; 3: spatial learning; 4: LTP; 5: NMDAR; 6: visual learning; 7: NMDAR1. Because these constraints are satisfiable, each constraint was arbitrarily assigned a weight of 1.

yielding—at each iteration, and for each experiment-selection policy—an equivalence class of consistent causal graphs. After each experiment, we recorded the number of graphs that remained in each equivalence class. This process continued until we performed every one of the 48 two-variable experiments defined for the research-map schema. This simulation was repeated for every one of the 543 possible DAGs over four variables, thus showing that the experiment-selection policies are not sensitive to specific features of the true causal graph, such as the density of its edges. For each policy, we then computed the average number of graphs in the equivalence class that remained after each iteration (Figure 8.10). Algorithm 4 provides pseudocode for this simulation.

The comparison of Algorithms 2 and 3 to random experiment selection does not imply that scientists are currently selecting their experiments at random. Instead, random experiment selection is used to establish a baseline of performance against which other methods can be judged; this

Data: G_A : all DAGs over N variables;
 P_A : all experiments over N variables and their results, for each DAG $G \in G_A$

Result: $S_{P,G}$: sequences of experiments;
 $S_{E,G}$: sequences of equivalence class sizes after each experiment

for each DAG $G \in G_A$ **do**

equivalence class $E \leftarrow G_A$;

set of performed experiments $P \leftarrow \emptyset$;

while $|P| < |P_{A,G}|$ **do**

$s \leftarrow$ experiment selected by policy (random, Alg. 2, or Alg. 3);

$P \leftarrow P \cup \{s\}$;

update E based on result of s for G (Chapter 5);

record s in $S_{P,G}$;

record $|E|$ in $S_{E,G}$;

end

end

compute average S_E across every DAG $G \in G_A$;

Algorithm 4: Simulation of experiment-selection policies. This simulation was performed once for each experiment-selection policy: (1) random, (2) Algorithm 2, and (3) Algorithm 3. This resulted in three sequences of average equivalence class sizes, which are displayed in Figure 8.10.

approach has precedent in the experiment-selection literature [VJM00, Vat01, KWJ04, VDB06]. Although scientists do not perform their experiments randomly, scientists usually do not plan their experiments in perfect coordination. These simulations thus highlight the experimental effort that can be saved when experiment planning is augmented by computational tools, and when research efforts in a field are coordinated [RFF15].

The results of these simulations illustrate a few key points about the limitations of piecemeal causal discovery and the importance of planning experiments in light of the causal explanations that remain viable. It is known that $\log(N) + 1$ experiments suffice to identify the true, causally sufficient DAG over N variables, where in each experiment, scientists can observe every variable, and intervene on any number of variables in the system. If we are limited to single-intervention

experiments, $N - 1$ experiments are sufficient and in the worst case necessary [EGS06, HEH13]. The context we consider here is further constrained: we consider studies in which only two variables are observed simultaneously and at most one variable can be intervened on per experiment. Thus, on average, between four and five graphs remain in the equivalence class after every possible two-variable experiment has been performed. Our policies' failure to uniquely identify some of the true causal graphs is in part a manifestation of the limits on piecemeal causal discovery [Ebe13, May13]. This underdetermination is also due in part to limitations of the research-map schema, including its lack of a conditioning set, which is discussed in § 9.5.2.

These simulations show that strategic experiment planning can save a considerable amount of effort in the laboratory: equivalent levels of underdetermination are reached with far fewer experiments using the suggestions of Algorithms 2 and 3. Compared to the policy of Algorithm 3, the random policy on average takes 32 additional studies to reach the minimum value. Algorithm 3 reaches an equivalence class of fewer than 10 graphs in less than half the number of experiments required by the random policy (9 vs. 19).

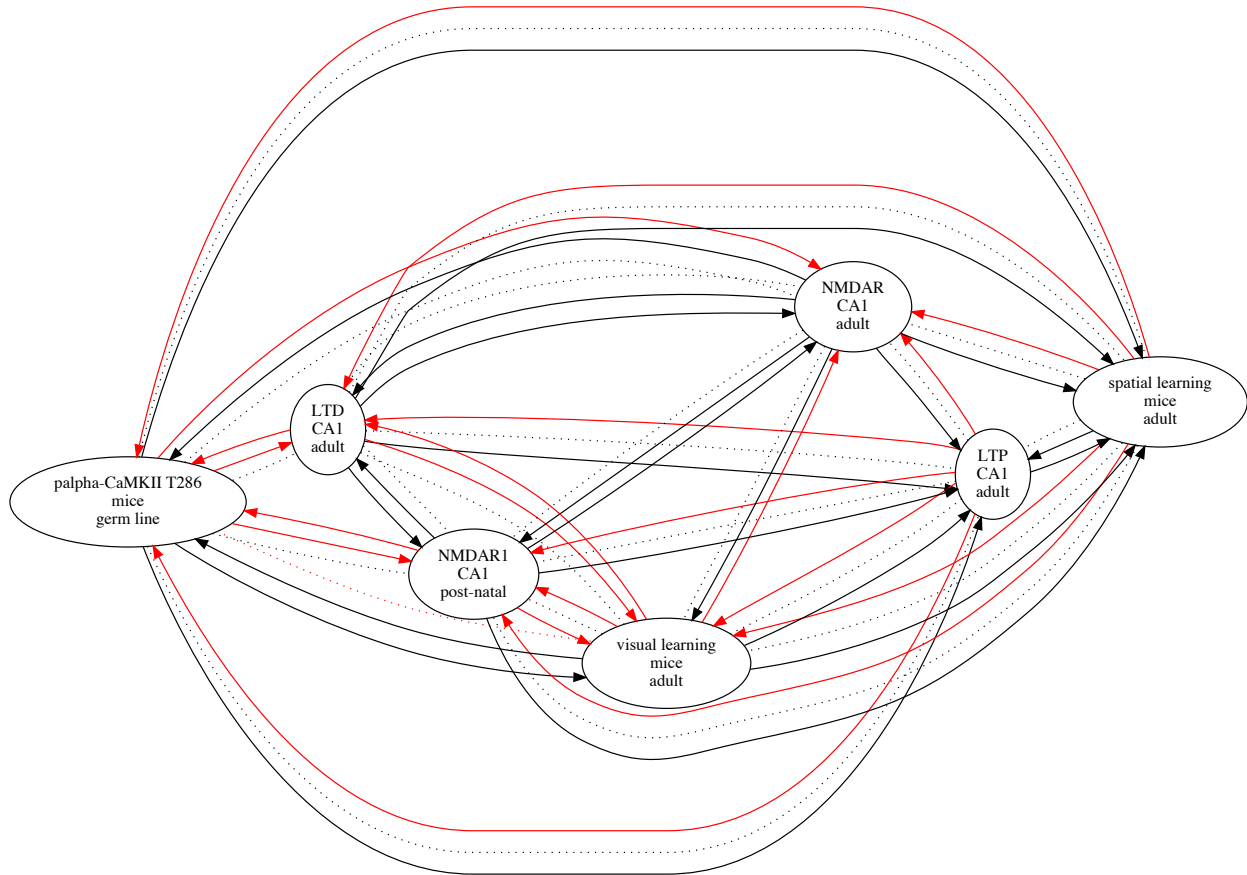


Figure 8.8: The “degrees of freedom” of an equivalence class. Each edge represents one of a causal graph’s *degrees of freedom*—i.e., one of the edge relations that can exist between two nodes in a causal graph. A dotted line denotes the relation in which the pair of nodes has no direct edge between them in the corresponding causal graph (e.g., $X \not\rightarrow Y$). Black edges are present in at least one causal graph in the equivalence class. Red edges are *not* present in the equivalence class, representing hypotheses that are inconsistent with the available evidence. The one red dotted edge—between “palpha-CaMKII T286” and “visual learning”—implies that every possible causal graph in the equivalence class has a direct edge between these nodes. This diagram demonstrates that even among the graphs that accommodate all the annotated constraints, many causal edges remain viable. Additional constraints—and thus additional experiments—are needed to eliminate edges from consideration. Note, however, that many edge relations have already been ruled out: the available evidence already precludes many edges (in red) from appearing in any of the consistent graphs. Such implications would be prohibitively difficult for a researcher to calculate by hand. © 2017 IEEE [MWW17b].

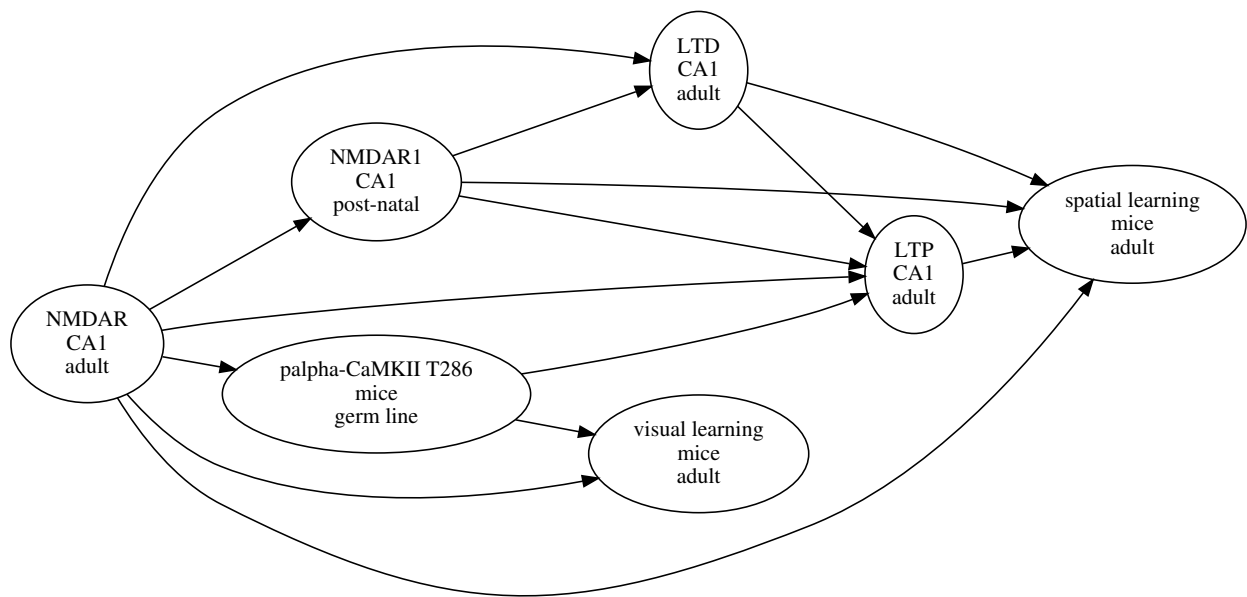


Figure 8.9: One of the thousands of optimal causal graphs derived from annotated results in literature. Each edge is a viable degree of freedom (Figure 8.8). © 2017 IEEE [MWW17b].

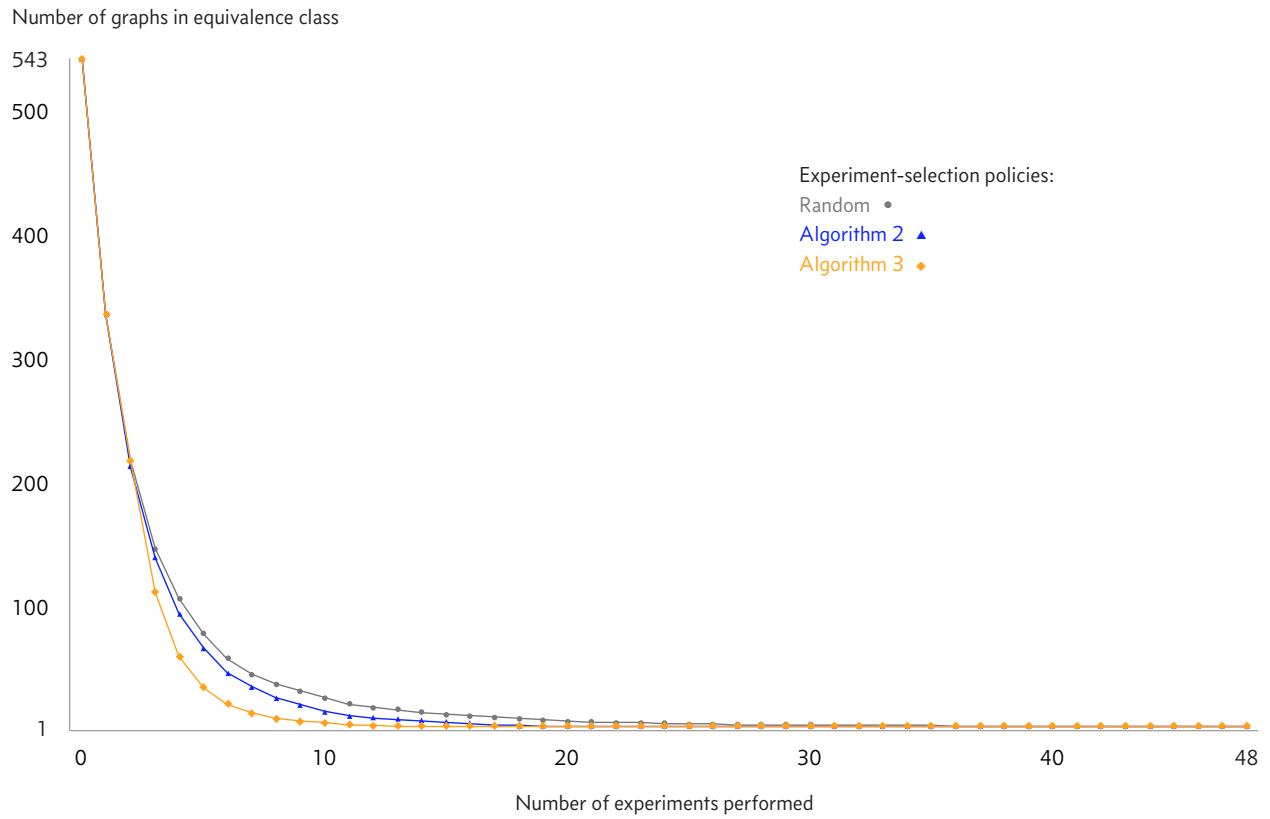


Figure 8.10: A comparison of three experiment-selection policies: (1) random, (2) Algorithm 2, and (3) Algorithm 3. This plot shows the results of the simulation given in Algorithm 4 for $N = 4$. The results show the experimental effort that is saved when each experiment is chosen based on the remaining degrees of freedom in the equivalence class.

CHAPTER 9

Conclusion

9.1 Summary of contributions

This dissertation presents methods that can help scientists to evaluate evidence and plan experiments. These research tasks are formalized using the graphical representations of research maps and causal graphs. With these representations, scientists can express their domain knowledge in computable data structures that can be used to automate parts of the scientific method, including evidence amalgamation, causal reasoning, and experiment planning.

Evidence amalgamation was formalized by redefining the research-map framework’s cumulative evidence index (CEI) using a Bayesian model. Although views are split on whether human learning is Bayesian [ED11], Bayesian models have been useful as an analytic basis for describing the learning process and have been supported by cognitive scientists [LYL08, TKG11, Gop12]. The CEI (§ 6.1) is intended not to resolve this debate but to offer a quantitative model of scientific consensus that captures scientists’ commonsense intuitions, many of which are currently applied only informally and qualitatively. The CEI’s primary strength is that it explicitly quantifies not only evidential consistency but also evidential convergence; the latter is an important concept that is commonly discussed in science but currently absent from traditional meta-analysis [MS18]. As a quantitative model of scientific consensus building, the CEI provides another analytic basis for meta-research. By mapping qualitative epistemic principles like convergence into quantitative parameters of a Bayesian model, the CEI advances the discussion of how scientists could evaluate evidence and justify future experiments more objectively. Together with traditional statistics like p -values, the CEI can thus offer a more holistic picture of the strength of evidence.

Meta-analytic causal reasoning was formalized by using research maps to express qualitative evidence from literature as formal constraints on causal structure; these constraints then drive a

causal discovery method that identifies causal explanations consistent with the evidence from literature. With this pipeline, scientists can synthesize research articles and plan experiments in light of *every* causal explanation that remains consistent with what is known, thus minimizing bias for specific causal structures. Biologists in particular may benefit from this approach: rather than simply stitching together pathway diagrams from the literature, the pipeline can be used to fuse these diagrams according to the rigorously defined formalisms of causal graphical models. The resulting causal graphs give logically consistent interpretations of the empirical evidence that motivated each individual pathway. Although the research-map schema was shown to be an effective representation for facilitating this process, we also identified multiple improvements that could be made to the schema to increase its expressivity, which in turn will allow for more efficient pruning of equivalence classes.

Experiment planning was operationalized by defining interpretable metrics for evidence and uncertainty, all of which are framed using graphical representations that are already familiar to scientists. The ResearchMaps web application and meta-analytic pipeline offer practical tools that leverage recent advances in causal discovery, thereby augmenting scientists' existing workflows. The use cases (§§ 8.1–8.2) and simulations (§ 8.3) involving these tools demonstrate their practical use, as well as their limitations—particularly as they pertain to the constraints that scientists face when designing experiments. This work thus demonstrates real examples of piecemeal causal discovery, which characterizes much of the work in biology [May13].

9.2 Assessment of hypotheses

This work shows that qualitative information in the literature is sufficient to drive causal discovery, yielding causal graphs that convey valid inferences and logically consistent explanations. Because the proposed methods use constraint-based causal discovery, multiple types of evidence can be integrated, including both primary data and statistical relations. This meta-analytic approach to evidence amalgamation and experiment selection can be readily applied by scientists, who must often deal with a mixture of evidence, much of which is expressed qualitatively in the literature.

This work also shows that evidence amalgamation and experiment selection can be partially automated and thus made less prone to error if scientists' domain knowledge is expressed using

computable data structures that can automate causal reasoning. By grounding these crucial research tasks in graphical representations of causality, scientists have an analytic basis for defending their interpretations, forming new hypotheses, and justifying their experimental designs.

9.3 Generalizability of the results

The meta-analytic method for causal discovery and experiment selection (Chapters 4–7) can be applied to any system with distinct and defined variables in which experimentation (or at least observation) is possible. This method is thus generalizable to most scientific domains. An important advantage of this method is that it can be applied in the absence of background knowledge: apart from labels (e.g., indices) to distinguish the variables, the pipeline does not need any information besides conditional (in)dependence relations—for instance, it does not need a domain-specific ontology—either to infer causal structures or to suggest additional experiments. This means that scientists can apply our method without first constructing a knowledgebase with relevant descriptions of the domain—a potentially expensive task that is required for some experiment-selection methods [KWJ04]. This feature is possible due to the universality of what we are trying to learn—namely, causal relations. In causal graphical models, the definition of a causal relation does not differ depending on the variables of interest: a causal graph involving biological phenomena is not fundamentally different from a causal graph involving economic phenomena, for instance.

Because it can operate without background knowledge, this method may suggest experiments that are infeasible—due perhaps to ethical concerns, or due to a lack of the requisite technology. But if a human can review the suggestions, infeasible experiments can simply be ignored; the researcher can instead traverse a list of ranked experiments until a feasible one is reached. And because technology is always changing, infeasible experiments may soon become feasible. It is therefore valuable to consider the potential information gain of experiments while ignoring whether they can currently be performed. By identifying which experiments would be most instructive, this type of analysis can thus help to prioritize the development of technology that would enable highly informative experiments.

9.4 Range of applicability

In this dissertation, experiment selection starts by assuming that the variables exist and that they are well defined; what is left is to determine the causal relations that describe the variables' interactions. But much of the work in science involves defining what these variables should be in the first place. This problem is not trivial; it requires cognitive processes that are beyond what we understand and can automate, including the *frame problem* [MH69]. Therefore, the methods in this dissertation cannot fully automate experiment selection in its most general sense.

However, these methods can be used to automate experiment selection for robots that are designed to perform many consecutive experiments on a predefined set of phenomena—a strategy that is increasingly common as scientific discovery is automated [Gly04, KCM18]. Examples include drug development [Mur11]; experiments to determine gene function [KWJ04]; and experiments to determine the effects of chemical compounds on the subcellular localization of proteins [NKS16]. In these contexts, scientists will usually not want to monitor thousands of consecutive experiments and select a new experiment at every step. The robot, left to its own devices, must therefore have some way to decide which experiments to perform, and in which order. This sequence of experiments could be completely defined by scientists beforehand. But ideally, the results of earlier experiments would inform the selection of later experiments, potentially saving time and resources. For these scenarios, the methods presented here offer heuristics with which the robot can choose an experiment sequence not just automatically but intelligently.

These methods should also be used to augment scientists' causal reasoning and experiment planning—cognitive skills that are being increasingly taxed as science becomes more complex. For some tasks, such “mind–machine partnerships” could help scientists more than even a human collaborator could [FBB18]—particularly in helping to avoid cognitive biases when considering causal explanations that are equally consistent with the evidence [MNB17].

These methods can help scientists to ensure that they tell logically consistent stories about sets of related experiments, as demonstrated in § 8.1. This is achieved by using a machine-readable data structure (e.g., a research map) to express evidence at the level of causal structure, and algorithmically checking its logical consistency. Growing databases of coded empirical results can be

continuously scanned for conflicts, alerting users whenever inconsistent results are entered. This approach can expand the scope of quality-assurance methods that are currently used on knowledge networks.

In addition to conflict detection, these methods can also help scientists to identify causal inferences that remain latent in the literature. Finding such inferences can be a challenge because it is often not obvious which results—and thus which research articles—should be considered together. For instance, § 8.2 gives a real example from the neuroscience literature of how we can make causal inferences involving phenomena that may have never been involved in the same experiment. This approach can be generalized: growing databases of machine-coded results can be continuously scanned for inferences using the methods presented here. When multiple constraints on causal structure are combined to derive an inference, we can check whether the constraints came from a single article or from multiple articles. In the latter case, scientists can be alerted to this advantageous grouping of articles.¹

It even seems feasible that as we improve methods to measure evidence and experiments' potential information gain, such measures could inform how research funding is allocated. This will likely not happen for some time, but it may be unhelpful to rule it out completely. It is worth remembering, for example, that some ecologists initially objected to statistics on the basis that it could not fully account for what made each individual organism distinct—a perspective that of course has fallen out of favor among scientists [GKN18]. Thus, if used properly, the types of methods presented in this dissertation could further democratize how funding is allocated.

9.5 Future work

The work in this dissertation can be improved in many ways. Below, a sample of such improvements are briefly discussed.

9.5.1 Automating literature annotation

The pipeline presented in Figure 1.4 would be much more scalable if we used natural language processing (NLP) to automate the annotation of research maps. This is likely to prove challenging:

¹ I am grateful to Frederick Eberhardt for pointing out this useful application.

the information needed to instantiate each research-map edge is often scattered throughout the research article, sometimes even in its figures and supplemental material. However, ResearchMaps already contains thousands of experiments annotated manually by domain experts; this information could in principle help to train classifiers to extract at least part of each research map automatically.² Noisy annotations could be corrected as needed by human reviewers. And partial annotations—for instance, the statement of an independence between two variables, without additional context—could still be translated into meaningful constraints on causal structure, thus pruning the set of viable models.

9.5.2 Extending the research-map schema

The research-map schema can be augmented to accommodate a greater variety of causal information. This would allow a greater variety of empirical results to be annotated and improve the ability of the annotated constraints to drive causal discovery.

One strategy for augmenting the research-map schema would be to simply adopt the schema for (in)dependence relations presented in Chapter 4 [HEJ14]. This would introduce a representation of statistical conditioning; this information is currently absent, so every research map annotation translates to a causal-structure constraint with an empty conditioning set C (Chapter 4). The schema would also be augmented to allow for any number of variables to be included in a study, all the way up to N , the number of variables in the system. This would greatly facilitate discovery, as it has been shown that if k , the number of variables observed simultaneously is less than N , there may always be some remaining underdetermination of the system’s causal structure [May13].

Another improvement is related to the asymmetry with which pairs of phenomena are annotated. Currently, each annotation names an agent and a target, with an edge assumed to be directed toward the target (agent \rightarrow target). This convention is used even for non-intervention experiments, where the agent and target are passively observed. Originally, this design decision was made to reflect the fact that in biology, studies are always designed in the context of other considerations, such as hypothetical connections and other intervention experiments. However, this assumption of directionality could be relaxed, enabling researchers to use the results of observational studies in

² I am grateful to Justin Wood for his preliminary work in this area.

which this directional information is not known or even hypothesized. For example, assuming not only that there is a causal relation between X and Y but also that X does not necessarily precede Y , passively observing both variables will reveal that they are correlated. This evidence is consistent with $X \rightarrow Y$, $X \leftarrow Y$, and indirect paths in both directions. The schema should thus leave open these possibilities.

9.5.3 Generalizing the cumulative evidence index in research maps

The cumulative evidence index (CEI) is defined for each edge in a research map; therefore, each index quantifies evidence and suggests experiments that involve only two nodes. The CEI would be more informative if its Bayesian model were generalized so that it could measure evidence and suggest experiments at the level of entire research maps with multiple edges. This dissertation gives qualitative principles that could guide the development of this generalized model (§ 7.1). The intention is for these qualitative principles to be formalized into a quantitative model for map-level integration, just as qualitative notions of causality were formalized with graphical models [Shi02].

9.5.4 Scaling SAT-based causal discovery methods

Although SAT-based causal discovery procedures achieve remarkably accurate results, they do not yet scale as well as other causal discovery algorithms. The pipeline presented in this dissertation would thus be made more useful to scientists if its causal discovery algorithm were made faster. One approach to improving the scalability of this method is to parallelize the algorithm. This parallelization could occur in two ways. First, a new ASP encoding (Chapter 5) could divide the overall SAT problem into separate problems, each of which could be solved simultaneously by separate solvers. Second, one could use a parallelized SAT solver [MML11a, MML11b, MML12b, MML12c, MML12a, HS18].³ It may also be possible to combine these two strategies.

Another strategy is to apply SAT-based causal discovery in highly constrained domains and explore the trade-offs that occur between accuracy and speed [Ebe17]. For instance, Magliacane, et al. [MCM16] significantly improved the scalability of the method by Hyttinen et al. [HEJ14] by considering only *ancestral* causal relations. Hyttinen et al. [HPJ16] considered sub-sampled time

³ Frederick Eberhardt pointed out this distinction between strategies.

series data and was able to scale their method to around 70 variables.

It is also worth noting that in many applications, scientists can wait long periods of time for hardware-optimized supercomputers to return a solution [Ebe17]. Compared to the enormous amounts of time that scientists spend reading the literature and planning experiments, waiting weeks or even months to identify *every* consistent causal explanation—among trillions, potentially—seems to be a relatively small price to pay.

9.5.5 Incorporating sign information into causal discovery

In the current pipeline (Figure 1.4), each research map’s sign information—whether each correlation is positive or negative—is lost in its translation to causal-structure constraints: every correlation, regardless of its sign, is mapped to a statistical dependence because the constraint-based causal discovery method does not use the sign information (Table 4.1). But this does not need to be the case. In principle, one could modify the algorithm to use this information in its search over causal structures, which may improve its performance.⁴

9.5.6 Improving experiment-selection heuristics

The experiment-selection policies presented in this dissertation were designed to be practical for scientists, meaning that they could be readily interpreted and incorporated into existing workflows. These heuristics could be made more efficient while still remaining applicable by practicing scientists. For instance, the analyses presented here considered only DAGs; however, feedback mechanisms are ubiquitous in biology, suggesting that these methods should be extended to allow for cycles, and thus the fourth degree of freedom discussed in § 6.2. One could also explore experiment-selection methods that are hybrids of the methods presented individually in § 7.1 and § 7.2. The intuition would be to choose experiments that improve the convergence and consistency of the evidence from a methodological perspective while also identifying the system’s true causal structure as efficiently as possible.

⁴ Frederick Eberhardt pointed out this strategy.

REFERENCES

- [ABB00] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, et al. “Gene Ontology: tool for the unification of biology.” *Nature Genetics*, **25**(1):25–29, 2000.
- [AKR17] Valentin Amrhein, Fränzi Korner-Nievergelt, and Tobias Roth. “The earth is flat ($p > 0.05$): significance thresholds and the crisis of unreplicable research.” *PeerJ*, **5**:e3544, 2017.
- [ALA11] Alexander V. Alekseyenko, Nikita I. Lytkin, Jizhou Ai, Bo Ding, Leonid Padyukov, Constantin F. Aliferis, and Alexander Statnikov. “Causal graph-based analysis of genome-wide association data in rheumatoid arthritis.” *Biology Direct*, **6**(1):25, 2011.
- [ASS18] David B. Allison, Richard M. Shiffrin, and Victoria Stodden. “Reproducibility of research: Issues and proposed remedies.” *Proceedings of the National Academy of Sciences*, 2018.
- [Bar03] Chitta Baral. *Knowledge Representation, Reasoning and Declarative Problem Solving*. Cambridge University Press, 2003.
- [BHM09] Armin Biere, Marijn Heule, and Hans van Maaren. “Handbook of satisfiability.” In *Frontiers in Artificial Intelligence and Applications*, volume 185. IOS Press, 2009.
- [Bod04] Olivier Bodenreider. “The unified medical language system (UMLS): integrating biomedical terminology.” *Nucleic Acids Research*, **32**(suppl 1):D267–D270, 2004.
- [BOH11] M. Bostock, V. Ogievetsky, and J. Heer. “D³: Data-driven documents.” *IEEE Transactions on Visualization and Computer Graphics*, **17**(12):2301–2309, 2011.
- [BP16] Elias Bareinboim and Judea Pearl. “Causal inference and the data-fusion problem.” *Proceedings of the National Academy of Sciences*, **113**(27):7345–7352, 2016.
- [CAS16] Denise J. Cai, Daniel Aharoni, Tristan Shuman, Justin Shobe, Jeremy Biane, Weilin Song, Brandon Wei, Michael Veshkini, Mimi La-Vu, Jerry Lou, et al. “A shared neural ensemble links distinct contextual memories encoded close in time.” *Nature*, **534**(7605):115, 2016.
- [CBB15] Gregory F. Cooper, Ivet Bahar, Michael J. Becich, Panayiotis V. Benos, Jeremy Berg, Jeremy U. Espino, Clark Glymour, Rebecca Crowley Jacobson, Michelle Kienholz, Adrian V. Lee, Xinghua Lu, Richard Scheines, et al. “The center for causal discovery of biomedical knowledge from big data.” *Journal of the American Medical Informatics Association*, **22**(6):1132–1136, 2015.
- [CES07] Lin S. Chen, Frank Emmert-Streib, and John D. Storey. “Harnessing naturally randomized transcription to infer regulatory relationships among genes.” *Genome Biology*, **8**(10):R219, 2007.

- [CFK02] Rui M. Costa, Nikolai B. Federov, Jeff H. Kogan, Geoffrey G. Murphy, Joel Stern, Masuo Ohno, Raju Kucherlapati, Tyler Jacks, and Alcino J. Silva. “Mechanism for the learning deficits in a mouse model of neurofibromatosis type 1.” *Nature*, **415**(6871):526–530, 2002.
- [CK07] Tim Clark and June Kinoshita. “Alzforum and SWAN: The present and future of scientific web communities.” *Briefings in Bioinformatics*, **8**(3):163–171, 2007.
- [CS03] Rui M. Costa and Alcino J. Silva. “Mouse models of neurofibromatosis type I: Bridging the GAP.” *Trends in Molecular Medicine*, **9**(1):19–23, 2003.
- [Dan05] David Danks. “Scientific coherence and the fusion of experimental results.” *The British Journal for the Philosophy of Science*, **56**(4):791–807, 2005.
- [Dar09] Adnan Darwiche. *Modeling and Reasoning with Bayesian Networks*. Cambridge University Press, 2009.
- [Don06] Kevin Donnelly. “SNOMED-CT: The advanced terminology and coding system for eHealth.” *Studies in Health Technology and Informatics*, **121**:279, 2006.
- [DP17] David Danks and Sergey Plis. “Amalgamating evidence of dynamics.” *Synthese*, pp. 1–18, 2017.
- [DQQ15] Fabricio H. Do-Monte, Kelvin Quiñones-Laracuente, and Gregory J. Quirk. “A temporal shift in the circuits mediating retrieval of fear memory.” *Nature*, **519**(7544):460, 2015.
- [Ebe05] Frederick Eberhardt. “On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables.” In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 178–184, Edinburgh, Scotland, 2005.
- [Ebe07] Frederick Eberhardt. *Causation and Intervention*. Ph.D. thesis, Carnegie Mellon University, 2007.
- [Ebe08] Frederick Eberhardt. “Almost optimal intervention sets for causal discovery.” In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 161–8, Helsinki, Finland, 2008.
- [Ebe09] Frederick Eberhardt. “Introduction to the epistemology of causation.” *Philosophy Compass*, **4**(6):913–925, 2009.
- [Ebe13] Frederick Eberhardt. “Experimental indistinguishability of causal structures.” *Philosophy of Science*, **80**(5):684–696, 2013.
- [Ebe16] Frederick Eberhardt. “Green and grue causal variables.” *Synthese*, **193**(4):1029–1046, 2016.
- [Ebe17] Frederick Eberhardt. “Introduction to the foundations of causal discovery.” *International Journal of Data Science and Analytics*, **3**(2):81–91, 2017.

- [ED11] Frederick Eberhardt and David Danks. “Confirmation in the cognitive sciences: The problematic case of Bayesian models.” *Minds and Machines*, **21**(3):389–410, 2011.
- [EGK01] John Ellson, Emden Gansner, Lefteris Koutsofios, Stephen C. North, and Gordon Woodhull. “Graphviz—open source graph drawing tools.” In *Proceedings of the International Symposium on Graph Drawing*, pp. 483–484, Berlin, 2001.
- [EGS06] Frederick Eberhardt, Clark Glymour, and Richard Scheines. “N–1 Experiments Suffice to Determine the Causal Relations Among N Variables.” In Dawn E. Holmes and Lakhmi C. Jain, editors, *Innovations in Machine Learning: Theory and Applications*, volume 194. Springer-Verlag, 2006.
- [Eva15] Michael Evans. *Measuring statistical evidence using relative belief*. CRC Press, 2015.
- [Eva16] Michael Evans. “Measuring statistical evidence using relative belief.” *Computational and Structural Biotechnology Journal*, **14**:91–96, 2016.
- [FBB18] Santo Fortunato, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, Alessandro Vespignani, Ludo Waltman, Dashun Wang, and Albert-László Barabási. “Science of science.” *Science*, **359**(6379), 2018.
- [Fed72] Valerii Vadimovich Fedorov. *Theory of Optimal Experiments*. Academic Press, Inc., New York, NY, 1972.
- [FPP18] Lisa Fourtune, Jérôme G. Prunier, Ivan Paz-Vinas, Géraldine Loot, Charlotte Veyssière, and Simon Blanchet. “Inferring causalities in landscape genetics: An extension of Wright’s causal modeling to distance matrices.” *The American Naturalist*, **191**(4):491–508, 2018.
- [Fry90] Morten Frydenberg. “The chain graph Markov property.” *Scandinavian Journal of Statistics*, **17**:333–353, 1990.
- [GAA08] Daniel Gardner, Huda Akil, Giorgio A. Ascoli, Douglas M. Bowden, William Bug, Duncan E. Donohue, David H. Goldberg, Bernice Grafstein, Jeffrey S. Grethe, Amarnath Gupta, et al. “The neuroscience information framework: A data and knowledge environment for neuroscience.” *Neuroinformatics*, **6**(3):149–160, 2008.
- [Gel13] Andrew Gelman. “Commentary: P values and statistical practice.” *Epidemiology*, **24**(1):69–72, 2013.
- [GFF98] Karl Peter Giese, Nikolai B. Fedorov, Robert K. Filipkowski, and Alcino J. Silva. “Autophosphorylation at Thr286 of the α calcium-calmodulin kinase II in LTP and learning.” *Science*, **279**(5352):870–873, 1998.
- [GKK11] Martin Gebser, Benjamin Kaufmann, Roland Kaminski, Max Ostrowski, Torsten Schaub, and Marius Schneider. “Potassco: The Potsdam answer set solving collection.” *AI Communications*, **24**(2):107–124, 2011.

- [GKK15] Martin Gebser, Roland Kaminski, Benjamin Kaufmann, Marius Lindauer, Max Ostrowski, Javier Romero, Torsten Schaub, and Sven Thiele. “Potassco User Guide.” Technical report, University of Potsdam, May 2015.
- [GKN18] Jessica Gurevitch, Julia Koricheva, Shinichi Nakagawa, and Gavin Stewart. “Meta-analysis and the science of research synthesis.” *Nature*, **555**(7695):175, 2018.
- [GL88] Michael Gelfond and Vladimir Lifschitz. “The stable model semantics for logic programming.” In *Proceedings of the Fifth International Conference and Symposium on Logic Programming*, pp. 1070–1080, 1988.
- [Gly04] Clark Glymour. “The automation of discovery.” *Daedalus*, **133**(1):69–77, 2004.
- [Goo50] I. J. Good. *Probability and Weighing of Evidence*. Griffon, London, 1950.
- [Goo60] I. J. Good. “Weight of evidence, corroboration, explanatory power, information and the utility of experiments.” *Journal of the Royal Statistical Society, Series B (Methodological)*, **22**(2):319–331, 1960.
- [Goo67] I. J. Good. “On the principle of total evidence.” *The British Journal for the Philosophy of Science*, **17**(4):319–321, 1967.
- [Gop12] Alison Gopnik. “Scientific thinking in young children: Theoretical advances, empirical research, and policy implications.” *Science*, **337**(6102):1623–1627, 2012.
- [GP13a] Sander Greenland and Charles Poole. “Living with P values: resurrecting a Bayesian perspective on frequentist statistics.” *Epidemiology*, **24**(1):62–68, 2013.
- [GP13b] Sander Greenland and Charles Poole. “Rejoinder: Living with statistics in observational research.” *Epidemiology*, **24**(1):73–78, 2013.
- [GSR16] Sander Greenland, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, and Douglas G. Altman. “Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations.” *European Journal of Epidemiology*, **31**(4):337–350, 2016.
- [GVP90] Dan Geiger, Thomas Verma, and Judea Pearl. “Identifying independence in Bayesian networks.” *Networks*, **20**(5):507–534, 1990.
- [HB12] Alain Hauser and Peter Bühlmann. “Two optimal strategies for active learning of causal models from interventions.” In *Proceedings of the 6th European Workshop on Probabilistic Graphical Models*, 2012.
- [HCV15] Lewis G. Halsey, Douglas Curran-Everett, Sarah L. Vowler, and Gordon B. Drummond. “The fickle P value generates irreproducible results.” *Nature Methods*, **12**(3):179, 2015.
- [HEH13] Antti Hyttinen, Frederick Eberhardt, and Patrik O. Hoyer. “Experiment selection for causal discovery.” *The Journal of Machine Learning Research*, **14**(1):3041–3071, 2013.

- [HEJ14] Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. “Constraint-based causal discovery: Conflict resolution with Answer Set Programming.” In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 340–349, Quebec City, Quebec, 2014.
- [Her18] Miguel A Hernán. “The C-word: Scientific euphemisms do not improve causal inference from observational data.” *American Journal of Public Health*, pp. e1–e4, 2018.
- [HG08] Yang-Bo He and Zhi Geng. “Active learning of causal networks with intervention experiments and optimal designs.” *Journal of Machine Learning Research*, **9**(11), 2008.
- [HHE13] Antti Hyttinen, Patrik O. Hoyer, Frederick Eberhardt, and Matti Järvisalo. “Discovering cyclic causal models with latent variables: A general SAT-based procedure.” In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI)*, Bellevue, Washington, 2013.
- [Hil65] Austin Bradford Hill. “The environment and disease: association or causation?” *Proceedings of the Royal Society of Medicine*, **58**(5):295–300, 1965.
- [HKY07] Jin-Hee Han, Steven A. Kushner, Adelaide P. Yiu, Christy J. Cole, Anna Matyina, Robert A. Brown, Rachael L. Neve, John F. Guzowski, Alcino J. Silva, and Sheena A. Josselyn. “Neuronal competition and selection during memory formation.” *Science*, **316**(5823):457–460, 2007.
- [HPJ16] Antti Hyttinen, Sergey Plis, Matti Järvisalo, Frederick Eberhardt, and David Danks. “Causal discovery from subsampled time series data by constraint optimization.” In *Proceedings of the International Conference on Probabilistic Graphical Models (PGM)*, 2016.
- [HS18] Youssef Hamadi and Lakhdar Sais. *Handbook of Parallel Constraint Reasoning*. Springer, 2018.
- [Hum03] David Hume. *A Treatise of Human Nature*. Dover Philosophical Classics, 2003.
- [Hum16] David Hume. “An enquiry concerning human understanding.” In *Seven Masterpieces of Philosophy*, pp. 191–284. Routledge, 2016.
- [IFD15] John P. A. Ioannidis, Daniele Fanelli, Debbie Drake Dunne, and Steven N. Goodman. “Meta-research: evaluation and improvement of research methods and practices.” *PLOS Biology*, **13**(10):e1002264, 2015.
- [III16] Robert D. O. Ness III. *Bayesian Causal Inference of Cell Signal Transduction from Proteomics Experiments*. Ph.D. thesis, Purdue University, 2016.
- [Joh03] Gary Johns. “How methodological diversity has improved our understanding of absenteeism from work.” *Human Resource Management Review*, **13**(2):157–184, 2003.

- [KCM18] Ross D King, Vlad Schuler Costa, Chris Mellingwood, and Larisa N Soldatova. “Automating Sciences: Philosophical and Social Dimensions.” *IEEE Technology and Society Magazine*, **37**(1):40–46, 2018.
- [Kos02] Jan TA Koster. “Marginalizing and conditioning in graphical models.” *Bernoulli*, pp. 817–840, 2002.
- [Kra17] Peter M. Krafft. *A Rational Choice Framework for Collective Behavior*. Ph.D. thesis, Massachusetts Institute of Technology, 2017.
- [KRO09] Ross D. King, Jem Rowland, Stephen G. Oliver, Michael Young, Wayne Aubrey, Emma Byrne, Maria Liakata, Magdalena Markham, Pinar Pir, Larisa N. Soldatova, et al. “The automation of science.” *Science*, **324**(5923):85–89, 2009.
- [KWJ04] Ross D. King, Kenneth E. Whelan, Ffion M. Jones, Philip G. K. Reiser, Christopher H. Bryant, Stephen H. Muggleton, Douglas B. Kell, and Stephen G. Oliver. “Functional genomic hypothesis generation and experimentation by a robot scientist.” *Nature*, **427**(6971):247–252, 2004.
- [Lee11] Christopher Lee. “Empirical information metrics for prediction power and experiment planning.” *Information*, **2**(1):17–40, 2011.
- [LHC17] Simon Lewin, Maggie Hendry, Jackie Chandler, Andrew D. Oxman, Susan Michie, Sasha Shepperd, Barnaby C. Reeves, Peter Tugwell, Karin Hannes, Eva A. Rehfuss, et al. “Assessing the complexity of interventions within systematic reviews: development, content and use of a new tool (iCAT_SR).” *BMC Medical Research Methodology*, **17**(1):76, 2017.
- [LHM09] Nicolas Le Novere, Michael Hucka, Huaiyu Mi, Stuart Moodie, Falk Schreiber, Anatoly Sorokin, Emek Demir, Katja Wegner, Mirit I. Aladjem, Sarala M. Wimalaratne, et al. “The systems biology graphical notation.” *Nature Biotechnology*, **27**(8):735–741, 2009.
- [LS13] Anthony Landreth and Alcino J. Silva. “The need for research maps to navigate published work and inform experiment planning.” *Neuron*, **79**(3):411–415, 2013.
- [LTD16] Debbie A. Lawlor, Kate Tilling, and George Davey Smith. “Triangulation in aetiological epidemiology.” *International Journal of Epidemiology*, **45**(6):1866–1886, 2016.
- [LYL08] Hongjing Lu, Alan L. Yuille, Mimi Liljeholm, Patricia W. Cheng, and Keith J. Holyoak. “Bayesian generic priors for causal learning.” *Psychological Review*, **115**(4):955, 2008.
- [May13] Conor Mayo-Wilson. “The limits of piecemeal causal inference.” *The British Journal for the Philosophy of Science*, **65**(2):213–249, 2013.
- [MCK10] Marloes H. Maathuis, Diego Colombo, Markus Kalisch, and Peter Bühlmann. “Predicting causal effects in large-scale systems from observational data.” *Nature Methods*, **7**(4):247–248, 2010.

- [MCM16] S. Magliacane, T. Claassen, and J. M. Mooij. “Ancestral causal inference.” In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, pp. 4466–4474, Barcelona, December 2016.
- [MD18] Daniel Malinsky and David Danks. “Causal discovery algorithms: A practical guide.” *Philosophy Compass*, **13**(1), 2018.
- [MDR18] Ani Movsisyan, Jane Dennis, Eva Rehfuess, Sean Grant, and Paul Montgomery. “Rating the quality of a body of evidence on the effectiveness of health and social interventions: a systematic review and mapping of evidence domains.” *Research Synthesis Methods*, pp. 1–19, 2018.
- [MH69] John McCarthy and Patrick J. Hayes. “Some philosophical problems from the standpoint of artificial intelligence.” In *Machine Intelligence*, pp. 463–502. Edinburgh University Press, 1969.
- [MLM06] Stijn Meganck, Philippe Leray, and Bernard Manderick. “Learning causal Bayesian networks from observations and experiments: A decision theoretic approach.” In *Proceedings of the International Conference on Modeling Decisions for Artificial Intelligence*, pp. 58–69. Springer, 2006.
- [MML05] Stijn Meganck, Bernard Manderick, and Philippe Leray. “A decision theoretic approach to learning Bayesian networks.” Technical report, Vrije Universiteit Brussels, 2005.
- [MML11a] Ruben Martins, Vasco Manquinho, and Inês Lynce. “Parallel search for Boolean optimization.” In *RCRA International Workshop on Experimental Evaluation of Algorithms for Solving Problems with Combinatorial Explosion*, 2011.
- [MML11b] Ruben Martins, Vasco M. Manquinho, and Inês Lynce. “Exploiting cardinality encodings in parallel maximum satisfiability.” In *IEEE 23rd International Conference on Tools with Artificial Intelligence, (ICTAI)*, pp. 313–320, 2011.
- [MML12a] Ruben Martins, Vasco Manquinho, and Inês Lynce. “Clause sharing in deterministic parallel maximum satisfiability.” In *RCRA International Workshop on Experimental Evaluation of Algorithms for Solving Problems with Combinatorial Explosion*, 2012.
- [MML12b] Ruben Martins, Vasco M. Manquinho, and Inês Lynce. “Clause sharing in parallel MaxSAT.” In *Learning and Intelligent Optimization: 6th International Conference (LION)*, pp. 455–460, 2012.
- [MML12c] Ruben Martins, Vasco M. Manquinho, and Inês Lynce. “Parallel search for maximum satisfiability.” *AI Communications*, **25**(2):75–95, 2012.
- [MNB17] Marcus R. Munafò, Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P. A. Ioannidis. “A manifesto for reproducible science.” *Nature Human Behaviour*, **1**(0021), 2017.

- [MS18] Marcus R. Munafò and George Davey Smith. “Repeating experiments is not enough.” *Nature*, **553**:399–401, 2018.
- [Mur01] Kevin P. Murphy. “Active learning of causal Bayes net structure.” Technical report, U.C. Berkeley, 2001.
- [Mur11] Robert F. Murphy. “An active role for machine learning in drug development.” *Nature Chemical Biology*, **7**(6):327–330, 2011.
- [MW15] Stephen L. Morgan and Christopher Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, New York, 2nd edition, 2015.
- [MWD18] Nicholas J. Matiasz*, Justin Wood*, Pranay Doshi*, William Speier, Barry Beckemeyer, Wei Wang, William Hsu, and Alcino J. Silva. “ResearchMaps.org for integrating and planning research.” *PLOS One*, **13**(5):e0195271, 2018.
- [MWW17a] Nicholas J. Matiasz, Justin Wood, Wei Wang, Alcino J. Silva, and William Hsu. “Computer-aided experiment planning toward causal discovery in neuroscience.” *Frontiers in Neuroinformatics*, **11**(12), 2017.
- [MWW17b] Nicholas J. Matiasz, Justin Wood, Wei Wang, Alcino J. Silva, and William Hsu. “Translating literature into causal graphs: Toward automated experiment selection.” In *Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 573–576, Nov. 2017.
- [NGL13] Jane Noyes, David Gough, Simon Lewin, Alain Mayhew, Susan Michie, Tomas Pantoja, Mark Petticrew, Kevin Pottie, Eva Rehfuss, Ian Shemilt, et al. “A research and development agenda for systematic reviews that ask complex questions about complex interventions.” *Journal of Clinical Epidemiology*, **66**(11):1262–1270, 2013.
- [NKS16] Armaghan W. Naik, Joshua D. Kangas, Devin P. Sullivan, and Robert F. Murphy. “Active machine learning-driven experimentation to determine compound effects on protein patterns.” *eLife*, **5**:e10047, 2016.
- [OG74] David Bridston Osteyee and Irving John Good. *Information, Weight of Evidence, the Singularity Between Probability Measures and Signal Detection*. Springer-Verlag, Berlin, 1974.
- [Org14] World Health Organization. *WHO Guidelines for Indoor Air Quality: Household Fuel Combustion*. World Health Organization, Geneva, Switzerland, 2014.
- [Pea09] Judea Pearl. *Causality*. Cambridge University Press, 2nd edition, 2009.
- [Pea17] Judea Pearl. “Theoretical impediments to machine learning with seven sparks from the causal revolution.” Technical Report R-475, University of California, Los Angeles, 2017.

- [Pea18] Judea Pearl. “Challenging the hegemony of randomized controlled trials: A commentary on Deaton and Cartwright.” *Social Science & Medicine*, 2018.
- [PM18] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.
- [Pop59] Karl Popper. *The Logic of Scientific Discovery*. Basic Books, 1959.
- [Reu13] Alexander Reutlinger. *A Theory of Causation in the Social and Biological Sciences*. Palgrave Macmillan, 2013.
- [RFF15] Andrey Rzhetsky, Jacob G. Foster, Ian T. Foster, and James A. Evans. “Choosing experiments to accelerate collective discovery.” *Proceedings of the National Academy of Sciences*, **112**(47):14569–14574, 2015.
- [Rob73] Robert W. Robinson. “Counting labeled acyclic digraphs.” In Frank Harary, editor, *New Directions in the Theory of Graphs*, pp. 239–273. Academic Press, New York, 1973.
- [RR56] Hans Reichenbach and Maria Reichenbach. *The Direction of Time*. University of California Press, 1956.
- [RRH11] Thomas A. Russ, Cartic Ramakrishnan, Eduard H. Hovy, Mihail Bota, and Gully A. P. C. Burns. “Knowledge engineering tools for reasoning with scientific observations and interpretations: a neural connectivity use case.” *BMC Bioinformatics*, **12**(351):1–15, 2011.
- [SBS18] Richard M. Shiffrin, Katy Börner, and Stephen M. Stigler. “Scientific progress despite irreproducibility: A seeming paradox.” *Proceedings of the National Academy of Sciences*, **115**(11):2632–2639, 2018.
- [Sch78] Gideon Schwarz. “Estimating the dimension of a model.” *The Annals of Statistics*, **6**(2):461–464, 1978.
- [SGS00] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition, 2000.
- [SHH06] Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. “A linear non-Gaussian acyclic model for causal discovery.” *The Journal of Machine Learning Research*, **7**:2003–2030, 2006.
- [Shi02] Bill Shipley. *Cause and Correlation in Biology: A User’s Guide to Path Analysis, Structural Equations and Causal Inference*. Cambridge University Press, 2nd edition, 2002.
- [Shi14] Shohei Shimizu. “LiNGAM: non-Gaussian methods for estimating causal structures.” *Behaviormetrika*, **41**(1):65–98, 2014.

- [SLB14] Alcino J. Silva, Anthony Landreth, and John Bickle. *Engineering the Next Revolution in Neuroscience: The New Science of Experiment Planning*. Oxford University Press, 2014.
- [SLY05] Eric E. Schadt, John Lamb, Xia Yang, Jun Zhu, Steve Edwards, Debraj GuhaThakurta, Solveig K. Sieberts, Stephanie Monks, Marc Reitman, Chunsheng Zhang, et al. “An integrative genomics approach to infer causal associations between gene expression and disease.” *Nature Genetics*, **37**(7):710, 2005.
- [SM15] Alcino J. Silva and Klaus-Robert Müller. “The need for novel informatics tools for integrating and planning research in molecular and cellular cognition.” *Learning & Memory*, **22**(9):494–498, 2015.
- [Smi81] Herman W. Smith. *Strategies of Social Research: The Methodological Imagination*. Prentice Hall, 1981.
- [SMS12] Daniel J. Stekhoven, Izabel Moraes, Gardar Sveinbjörnsson, Lars Hennig, Marloes H. Maathuis, and Peter Bühlmann. “Causal stability ranking.” *Bioinformatics*, **28**(21):2819–2823, 2012.
- [SPP05] Karen Sachs, Omar Perez, Dana Pe’er, Douglas A. Lauffenburger, and Garry P. Nolan. “Causal protein-signaling networks derived from multiparameter single-cell data.” *Science*, **308**(5721):523–529, 2005.
- [SS04] Peter Spirtes and Richard Scheines. “Causal inference of ambiguous manipulations.” *Philosophy of Science*, **71**(5):833–845, 2004.
- [Stu98] Milan Studený. “Bayesian networks from the point of view of chain graphs.” In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 496–503. Morgan Kaufmann Publishers Inc., 1998.
- [SW00] Terri A. Scandura and Ethlyn A. Williams. “Research methodology in management: Current practices, trends, and implications for future research.” *Academy of Management Journal*, **43**(6):1248–1264, 2000.
- [THT96] Joe Z. Tsien, Patricio T. Huerta, and Susumu Tonegawa. “The essential role of hippocampal CA1 NMDA receptor-dependent synaptic plasticity in spatial memory.” *Cell*, **87**(7):1327–1338, 1996.
- [TK01] Simon Tong and Daphne Koller. “Active learning for structure in Bayesian networks.” In *Proceedings of the International Joint Conference on Artificial Intelligence*, volume 17, pp. 863–869. Lawrence Erlbaum Associates LTD, 2001.
- [TKG11] Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. “How to grow a mind: Statistics, structure, and abstraction.” *Science*, **331**(6022):1279–1285, 2011.
- [TT11] Francis-Noël Thomas and Mark Turner. *Clear and Simple as the Truth: Writing Classic Prose*. Princeton University Press, 2nd edition, 2011.

- [Vat01] Ivayla Nedeltcheva Vatcheva. *Computer-supported Experiment Selection for Model Discrimination*. Ph.D. thesis, University of Twente, Netherlands, 2001.
- [VC18] Veronica J. Vieland and Hasok Chang. “No evidence amalgamation without evidence measurement.” *Synthese*, pp. 1–23, 2018.
- [VDB06] Ivayla Vatcheva, Hidde De Jong, Olivier Bernard, and Nicolaas J. I. Mars. “Experiment selection for the discrimination of semi-quantitative models of dynamical systems.” *Artificial Intelligence*, **170**(4–5):472–506, 2006.
- [VDH13] Veronica J. Vieland, Jayajit Das, Susan E. Hodge, and Sang-Cheol Seok. “Measurement of statistical evidence on an absolute scale following thermodynamic principles.” *Theory in Biosciences*, **132**(3):181–194, 2013.
- [VH11] Veronica J. Vieland and Susan E. Hodge. “Measurement of evidence and evidence of measurement.” *Statistical Applications in Genetics and Molecular Biology*, **10**(1), 2011.
- [Vie06] Veronica J. Vieland. “Thermometers: Something for statistical geneticists to think about.” *Human Heredity*, **61**(3):144–156, 2006.
- [Vie11] Veronica J. Vieland. “Where’s the evidence?” *Human Heredity*, **71**(1):59–66, 2011.
- [VJM00] Ivayla Vatcheva, Hidde de Jong, and Nicolaas J. I. Mars. “Selection of perturbation experiments for model discrimination.” In *Proceedings of the 14th European Conference on Artificial Intelligence*, pp. 191–195. IOS Press, 2000.
- [VP91] Thomas S. Verma and Judea Pearl. “Equivalence and synthesis of causal models.” In P. P. Bonissone, M. Henrion, L. N. Kanal, and J. F. Lemmer, editors, *Uncertainty in Artificial Intelligence 6*, pp. 255–268. Elsevier, Amsterdam, 1991.
- [VS15] Veronica J. Vieland and Sang-Cheol Seok. “Statistical evidence measured on a properly calibrated scale across nested and non-nested hypothesis comparisons.” *Entropy*, **17**(8):5333–5352, 2015.
- [VS16] Veronica J. Vieland and Sang-Cheol Seok. “Statistical evidence measured on a properly calibrated scale for multinomial hypothesis comparisons.” *Entropy*, **18**(4):114, 2016.
- [WCS66] Eugene J. Webb, Donald Thomas Campbell, Richard D. Schwartz, and Lee Sechrest. *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. Rand McNally, Chicago, 1966.
- [Woo05] James Woodward. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, 2005.
- [Wri21] Sewall Wright. “Correlation and causation.” *Journal of Agricultural Research*, **20**(7):557–585, 1921.

- [Wri23] Sewall Wright. "The theory of path coefficients: A reply to Niles's criticism." *Genetics*, 8(3):239–255, 1923.
- [Wri34] Sewall Wright. "The method of path coefficients." *The Annals of Mathematical Statistics*, 5(3):161–215, 1934.
- [Zol10] Kevin J. S. Zollman. "The epistemic benefit of transient diversity." *Erkenntnis*, 72(1):17, 2010.